USAID
FROM THE AMERICAN PEOPLE

EARLY GRADE READING ASSESSMENT BASELINE REPORT

PUNJAB PROVINCE

SEPTEMBER 2014

This publication was produced for review by the United States Agency for International Development by. It was prepared by Management Systems International (MSI) with School-to-School International (STS) under the Monitoring and Evaluation Program (MEP).

EARLY GRADE READING ASSESSMENT BASELINE REPORT

PUNJAB

Contracted under Order No. AID-391-C-13-00005

Monitoring and Evaluation Program (MEP)

ACKNOWLEDGEMENTS

We would like to thank the Education team of USAID/Pakistan for their forward planning to be able to collect baseline data before the roll-out of the two important reading programs. Their support and responsiveness under a demanding timeline made this study possible. We would also like to thank the Department of School Education, Directorate of Primary Education, and the Government of Punjab for their support of this activity. Finally, this effort would not have been possible without the dedication of our field teams of quality control officers and our local data collection partner, the Society for the Advancement of Education (SAHE).

CONTENTS

List of Tables and Figures

ACRONYMS

AJK	Azad Jammu and Kashmir
B.A.	Bachelor of Arts
BEFARe	Basic Education for Awareness, Reforms and Empowerment
B.Sc.	Bachelor of Science
C.T.	Certificate of Teaching (Grade 12 plus FA/FSC Certificate)
DOE	Directorate of Education
EGRA	Early Grade Reading Assessment
F.A.	Fellow in Arts
FATA	Federally Administered Tribal Areas
F.Sc.	Fellow in Sciences
GB	Gilgit-Baltistan
ICT	Islamabad Capital Territory
KP	Khyber Pakhtunkhwa
M.A.	Master of Arts
Matric	Secondary School (Grade 10) Certificate (Matriculation)
M.Ed.	Master of Education
M.Sc.	Master of Science
MSI	Management Systems International
MT	Master Trainers
NEAS	National Education Assessment System
NEMIS	National Education Management Information System
PRP	Pakistan Reading Project
PTA	Parent Teacher Association
PTC	Parent Teacher Council
P.T.C.	Primary Teaching (Grade 12) Certificate
PTSMC	Parent Teacher School Management Committee
QCO	Quality Control Officer
SAHE	Society for the Advancement of Education
SPSS	Statistical Package for the Social Sciences
SRP	Sindh Reading Project
STS	School-to-School International
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Overview

In 2013, Management Systems International (MSI) and School-to-School International (STS) conducted a baseline reading assessment for primary school children prior to the launching of two USAID-funded projects: the Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP). PRP is targeting improved reading for 910,000 children in Azad Jammu and Kashmir (AJK), Balochistan, the Federally Administered Tribal Areas (FATA), Gilgit-Baltistan (GB), the Islamabad Capital Territory (ICT), Khyber Pakhtunkhwa (KP), and Sindh, while the SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and administrators, and 2) improved reading instruction for children in the primary grades.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. This report covers the baseline assessment in Punjab, which took place in October 2013 and in January 2014. In May 2013, GB, AJK, and ICT were part of Round 1 of the baseline data collection; Round 2 in KP and Sindh was completed in September 2013; and Round 3 in Balochistan, FATA, and Punjab was completed in October 2013 and January 2014. The following activities were carried out for all of the provinces, including Punjab: 1) design, 2) sampling, 3) instrumentation, 4) planning, 5) training, 6) implementation, 7) analysis, and 8) reporting.

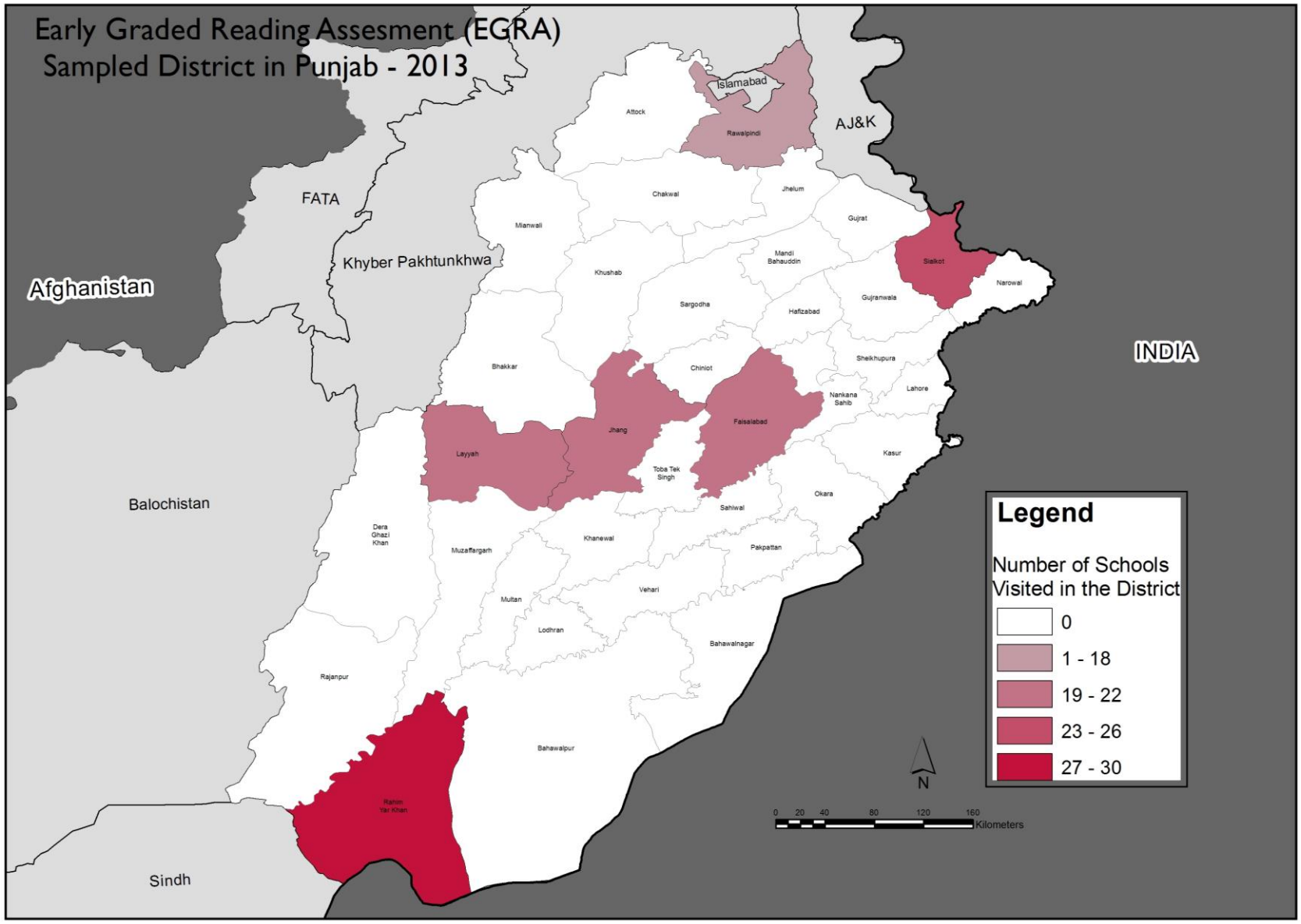
The external evaluation design, which was developed prior to the baseline assessment, was tailored to the implementation of the PRP and SRP in each province. In most provinces, Punjab included, a quasi-experimental design will be used, with two treatment groups: full treatment and light treatment. The full treatment group will receive both the first and second kinds of support, i.e., 1) policy, laws, and guidelines, and 2) improved instruction. The light treatment group will only receive the first kind of support.

In accordance with the USAID evaluation guidelines, students at two selected grade levels – grades 3 and 5 – will be assessed at three time points: baseline, midline, and endline. An internationally accepted assessment tool, the Early Grade Reading Assessment (EGRA), will be individually administered to over 33,000 children in 1,120 schools throughout the country. Over the course of the projects, the evaluators will compare the baseline results with those at the midline and endline to examine success in improving children's reading levels in Pakistan. The sampling was designed so that each province could be evaluated independently.

The long-term goal of this evaluation is to compare each province's baseline results to its midline and endline results, rather than other province's results. There are too many confounding variables – languages, curricula, administration dates, etc., that could render province-to-province comparisons meaningless. In addition, the reading experts who developed the core EGRA tool recommend against comparisons across languages due to differing structures. Furthermore, the evaluation is designed to investigate reading performance of the full and light treatment groups across time: baseline, midline, and endline. The differences between treatments will be fully investigated later using the baseline data as the starting point for comparisons. In-depth comparisons between the full and light treatment groups are not useful at this time; such comparisons at baseline could add some bias by facilitating competition between the two groups that could compromise the validity of the evaluation.

For the baseline in Punjab, all activities were completed by the end of January 2014. The EGRA baseline results were presented and discussed at a consultative meeting in April 2014. Representatives from the Department of School Education, Directorate of Primary Education (DOE), USAID, and the contractors (MSI and STS) attended the consultation. Revisions have been made to this report based on the discussions between the stakeholders.

Map of Sampled Districts



Key Points

Several key points from the EGRA baseline assessment in Punjab are highlighted below:

Implementation

1. The Punjab evaluation involves two kinds of comparisons: 1) a comparison of full and light treatment groups to determine the effects of full treatment above and beyond that of the light treatment, and 2) a comparison of each group to itself at the baseline, midline, and endline. Given the long-term design of this evaluation, this baseline report will not statistically test the differences between the groups' initial reading performance because doing so may confound the study by facilitating competition between the groups. The report will, however, present the baseline scores for each group. (Please see Figure 1 and the accompanying text for a fuller description of the evaluation design.)

2. District selection into full and light treatment groups was finalized following consultative meetings between the DOE and USAID in February 2013.

3. EGRA in Urdu was used in the Punjab province. The EGRA tools, which have been administered in various forms in over 40 countries, were successfully adapted for use in Pakistan. These included individually administered reading tests for students, along with questionnaires for students, teachers, and head teachers. The Urdu version of the tools was piloted in AJK, ICT, and KP. The Sindhi version of the tools was piloted in Sindh.

4. A total of 140 schools, with 70 schools from each group (full and light treatment), were selected for the baseline.

5. There were a total of 36 districts in Punjab, out of which 13 were full treatment and the remaining 23 were light treatment. A simple random sample of three districts – Rahimyar Khan, Faisalabad, and Rawalpindi – was chosen for full treatment. Three districts, Sialkot, Jhang, and Layyah, were selected for light treatment. The number of schools in the districts and the apportioned number of samples from each district by gender is shown in Table 2.

6. The EGRA testing window for Rahimyar Kahn, Faisalabad, Rawalpindi, Jhang, and Sialkot was October 2013, which was seven months (including a three-month holiday in the summer) after the start of the academic year. Due to some sampling issues that needed to be corrected, testing in Layyah was conducted in late January 2014, which was 10 months after the start of school (also including a three-month holiday in the summer). Because data in Layyah were collected three months after that of the other districts, the baseline scores may be slightly higher than they would have been if the data collection had been in October 2013. However, the EGRA scores for Layyah were similar to those of the other districts, and, despite this slight disparity in baseline data collection, we recommend maintaining a consistent testing calendar (October for all districts) at midline and endline and negating district comparisons.

7. For each district, a random sample of boys and girls schools was obtained, followed by a random sample of students in grades 3 and 5 within those schools. The results from this sample are presented in this report as a generalized view of the reading levels for students in the Punjab schools.

8. Please note that district comparisons are not possible because the districts were not evenly sampled; the number of sampled schools varied by district and the sample sizes are limited for each district. Moreover, the gap between the start of the school year and the EGRA administration fluctuated by district, thereby altering the amount of instructional time students received and potentially affecting the reading performance levels students achieved across the districts.

9. The target baseline sample for Punjab was 140 schools. The assessment tools were successfully administered in (with a percentage of the target reached in parentheses) 140 schools (100.0

percent) to 4,148 students (98.8 percent), 227 teachers (81.1 percent), and 140 head teachers (100.0 percent). When analyzing EGRA scores by gender, the sample size decreases to 4,136 due to 12 students with missing gender codes. The percent of teachers is relatively low because some teachers taught both grades and others did not indicate a grade. These responses were not counted in the survey results, which were analyzed by grade level.

10. The validity and reliability of the tools for both languages was acceptable. Validity was assured through the adaptation process, which involved 17 educationists from throughout the country who participated in a workshop in Islamabad. Reliability was assured through the high quality of the assessment tools and the standardized administration of the tools in Punjab. Reliability estimates (of internal consistency) were calculated using the coefficient alpha.

11. The data entry and data cleaning process followed international standards. All student data were entered twice into two separate databases. All data were reconciled across the two databases and with the assessment booklets. A clean data file was produced for analysis.

12. In the analysis phase, scores were calculated in three ways: 1) percentage correct scores for the reading tasks, 2) average percentage correct (grand means) for reading summary scores, and 3) adjusted raw scores for the timed tasks. These scores provide a comprehensive picture of student performance. Analysis of student, teacher, head teacher, and school characteristics was carried out using the summary scores.

Results

1. The EGRA was administered to 2,079 grade 3 students and 2,069 grade 5 students. The reliability was excellent for both grades (alpha = 0.85 for grade 3 and 0.84 for grade 5). These high reliabilities indicate that the items worked well in measuring reading constructs at both grades.

2. The task and item statistics showed that EGRA is able to discriminate between low- and high-achieving students in both grades. The task p-values for grade 3 provided a spread on the lower to lower-middle section of the difficulty range, while p-values for grade 5 were higher and covered the upper-lower half to the high-middle parts of the spectrum. All task scores at grades 3 and 5 had item-total correlations equal to or greater than 0.25, indicating good discrimination quality for these tasks. (Complete item statistics are listed in Annex 1.)

3. Grade 3 posted the highest scores in passage reading, familiar word reading, letter name recognition, and orientation to print, however, the percent correct scores were all below 50 percent. The most difficult tasks for the students in Punjab were comprehension (passage and listening) and phonics (non-word reading, letter sound knowledge, and phonemic awareness). The grade 5 students' scores showed similar patterns. The highest scores for grade 5 were in familiar word reading and passage reading; whereas the most challenging tasks were comprehension (passage and listening) and letter sound knowledge. The most challenging reading tasks for the students in Punjab were reading comprehension and phonics, particularly letter sound knowledge.

4. There was also substantial progression from grade 3 to grade 5 on the summary score (17.6 points). The phonics tasks of letter sound knowledge (9.0 points) and phonemic awareness (11.5 points) showed the lowest improvement. In contrast, the greatest gains were in familiar word reading (34.5 points), passage reading (31.5 points), passage comprehension (28.3 points), and non-word reading (27.0 points). In areas where there are large differences, interventions at grade 3, or in earlier grades, could have particularly large effects in accelerating children's learning. This progress was consistent across gender and treatment groups.

5. Boys and girls showed different patterns in reading skills. At grade 3, boys performed best on orientation to print (42.0 percent), letter name recognition (37.6 percent), and passage reading (35.7 percent). Grade 3 girls displayed higher scores in familiar word reading (50.2 percent) and passage reading (53.7 percent), followed by letter name recognition (48.0 percent), and orientation to print (39.5 percent). At grade 5, boys and girls were best at familiar word reading (69.9 percent) and passage reading (71.4 percent), but boys had difficulty with passage

comprehension (33.0 percent) and letter sound knowledge (18 percent), while girls were mostly challenged by letter sound knowledge (21.2 percent). In comparing scores between the genders, the girls' scores were significantly higher ($p < 0.001$) on the EGRA summary score and all tasks except orientation to print at grade 5 (higher, but not significantly). Again, the data show that students in Punjab have the most difficulty with letter sound knowledge and comprehension.

6. Students were timed on five tasks as they read letters, words, or passages. These tasks were categorized into phonics (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Students at both grades had lower phonics scores than reading-rate fluency scores. Moreover, gains from grade 3 to grade 5 were lower for phonics than for reading-rate fluency tasks. Girls' fluency rates were significantly higher than the boys for all tasks ($p < 0.001$). There were only small differences between the light and full treatment groups' fluency scores. Although the passage was designed for grade 3, this difference shows that the reading-rate fluency levels in grade 3 are low, but that students can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn. Specifically, mastery of phonics and phonemic awareness should help the students become better overall readers. It is clear that this type of knowledge and these skills are not receiving an appropriate emphasis in schools in Punjab.

7. Of the nine task percent correct scores, the full treatment group had higher scores on eight tasks for both grades, but these differences were small. This slight discrepancy will be corrected statistically at midline and endline by analyzing the growth for each group from baseline to midline and endline. Because this is a baseline report, treatment group differences are not statistically tested at this time.

8. In general, at the baseline, having someone read to children at home, having children read to someone else at home, and children reading silently at home were not related to higher reading scores. The only exception was grade 5 students who read silently at home had significantly ($p < 0.05$) higher scores than their peers who did not read silently at home.

9. Teacher and head teacher questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of variation in the scores that were related to their characteristics. For example, an analysis of student scores by teacher and head teacher education, certification, age, experience, and attendance at in-service trainings found no consistent patterns relating to lower or higher student scores.

10. For the schools, the few urban schools had higher reading scores than those in rural settings, however, this finding should be interpreted with caution due to the low number of urban schools in the sample. Girls and mixed-gender schools performed better at grade 3, while the boys and mixed-gender schools posted higher scores at grade 5. Over 90 percent of the schools reported having a Parent Teacher Association (PTA) and approximately 50 percent stated they had a school library. Neither of these factors, however, was related to higher reading scores. Lastly, 87 percent of the schools reported having better infrastructure and also had higher student reading scores; scores increased with the addition of water, electricity, or toilets.

Evaluation Recommendations

Given the success of the baseline assessment in Punjab (and in the other provinces), the methods used in 2013 should be repeated as much as possible for the midline and endline assessments in future years. It should be conducted as follows:

1. The EGRA instruments proved to be of high quality, and equivalent versions of those tools should be developed – through trans-adaptation, piloting, and revision – for the midline and endline assessments so that progress can be accurately measured over time.

2. The EGRA items and tasks had good discrimination (quality) values and covered the low-to-middle part of the difficulty range. At baseline, the reading scores were relatively low for both grades and show room for growth. In addition, histograms and box plots provided evidence that

the tool is expected to measure higher levels of reading-rate fluency that are anticipated following project-led interventions. Therefore, the baseline data indicates that EGRA is appropriate for measuring increases in reading ability at midline and endline.

3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. The data collection procedures were very effective, as 99 percent of the target was tested. They should be maintained in the midline and endline, i.e., keep the same districts and schools, along with the sampling methods at the school level.

4. Due to a sampling issue, one light treatment district, Layyah, was tested three months after the others. While this may have slightly increased the scores in Layyah at baseline and subsequently may decrease the gains at midline, those small discrepancies will not invalidate the endline results. Therefore, pending discussions with the USAID Lahore team, we recommend that future data collection in all of Punjab's sample districts should occur in October.

5. The systems for field data collection should be replicated, with the same systems for recruitment and training for the master trainers (MTs), field supervisors, quality control officers (QCOs), and enumerators as used in the baseline.

6. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.

7. The analysis should follow the same procedures, with calculations of reliability, difficulty, task percent-correct scores, summary scores, and timed task scores. The baseline, midline, and endline scores should be computed using the same procedures so that improvements in students' reading can be accurately examined over time.

8. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct for each task, summary score as well as fluency rates, are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.

9. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires to collect data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

CHAPTER 1: INTRODUCTION

The Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP) are two five-year initiatives funded by USAID. The projects/programs will cover over 40,000 government schools in Pakistan's eight provinces/areas/territories (hereafter referred to as provinces). PRP is targeting improved reading for 910,000 children in AJK, KP, FATA, GB, ICT, Balochistan, and Sindh, while the SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and educational administrators, and 2) improved reading instruction for children in primary grades. Some districts in Pakistan will receive both kinds of support, i.e., "full treatment," while others will receive only the policy support, i.e., "light treatment." All schools within districts will receive the same type of treatment.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. The evaluation baseline took place in 2013, prior to the launch of the reading interventions. In accordance with USAID program evaluation guidelines, samples of students in two selected grade levels – grade 3 and grade 5 – were assessed throughout Pakistan so that independent baselines can be established in each province. Students at the same grade levels will be assessed at the midline and endline time points to evaluate the success of the interventions, taking into account the two treatment groups. The goal of the evaluation is to conduct a long-term assessment for both treatment groups in each province.

This report covers Punjab province. Along with FATA and Balochistan, Punjab was part of Round 3 of the baseline data collection in October 2013 (and January 2014 for the Layyah District); data from Pakistan's other five provinces were collected in May 2013 (ICT, AJK, GB) and September 2013 (Sindh, KP). The following activities were implemented for all of the provinces, including Punjab:

1. Design – USAID required a cross-sectional design, i.e., assessing students at the same grade levels (grades 3 and 5) over the course of PRP and SRP. In most provinces, including Punjab, this was complemented by a quasi-experimental design with the two treatment groups (full and light).
2. Sampling – There were a total of 36 districts in Punjab, out of which 13 were full treatment and the remaining 23 were light treatment. A simple random sample of three districts for full treatment was chosen (Rahimyar Khan, Faisalabad, and Rawalpindi) and three districts were randomly selected for light treatment (Sialkot, Jhang, and Layyah). Schools were then apportioned according to location and gender. Balance for the location variable was not possible due to too few urban schools (8) sampled. Therefore, it was not appropriate to fully investigate the EGRA differences between urban and rural schools in Punjab.
3. Instrumentation – EGRA tools were developed, with tests at the grade 3-level in English, Sindhi, and Urdu, and questionnaires for teachers, head teachers, and students in Urdu and Sindhi. Model EGRA instruments were trans-adapted, piloted, revised, and finalized for use in Pakistan. The Urdu instruments were used in Punjab.
4. Planning – A field administration plan was developed for the baseline administration that would ensure the reliability of the data collected. The plan specified the timeline, training, logistics, field activities, supervision, data entry, analysis, reporting, and quality control.
5. Training – Workshops were conducted to train all MTs, supervisors, enumerators, and QCOs. Enumerators and supervisors were observed to ensure clear comprehension and skills adequate to implement the EGRA tools.
6. Implementation – The baseline survey was implemented according to the plan. It ensured that all of the field activities took place in a standardized manner, as verified by the QCOs. The fieldwork was followed by data entry and preparation of a clean data file.

7. Analysis – Data were analyzed using spreadsheet (Excel) and statistical (SPSS) software. Experienced statisticians/psychometricians conducted the analysis, produced data tables and graphs, and ensured quality control.

8. Reporting – Provincial-level reports were produced and will be disseminated to the provincial education authorities. A template was developed according to guidelines from the USAID contract.

This report is organized into four chapters: 1) introduction, 2) methodology, 3) findings and results, and 4) conclusions and recommendations. Annexes with item statistics, box plots for the timed tasks, and a possible process for establishing a reading proficiency threshold follow the chapters.

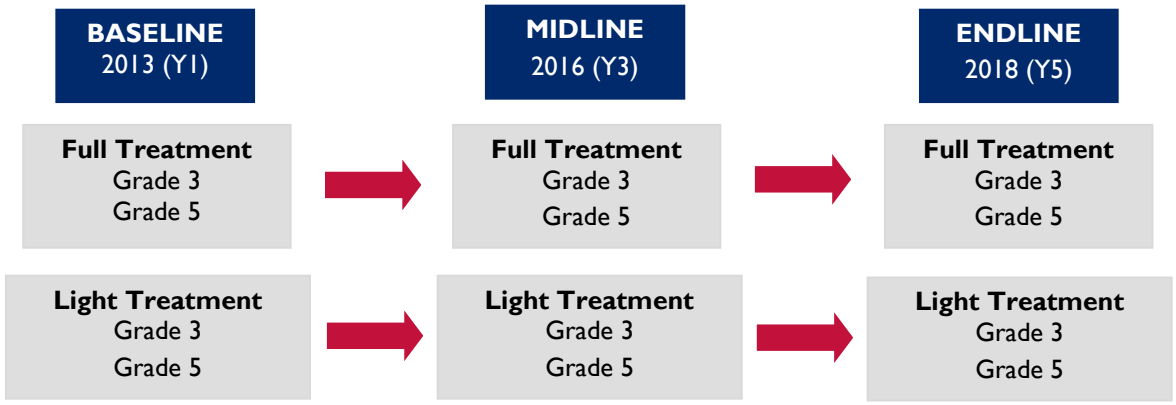
CHAPTER 2: DESIGN AND METHODOLOGY

This chapter presents the evaluation design and methodology, including the systems used for collecting the EGRA baseline data for schools in Punjab. There are sections on the evaluation design, timeline, sampling, instrument development, data collection, data entry, and data analysis.

Evaluation Design

Following USAID policy, a cross-sectional evaluation design was developed prior to the baseline data collection. As shown in Figure 1, the design features two grade levels (3 and 5) and three time points (baseline, midline, and endline). Different groups of grade 3 and grade 5 students will be compared against each other across the three time points. In the figure, the years for the midline and endline are approximate and may be altered in accordance with implementation of the PRP interventions.

FIGURE 1: EVALUATION DESIGN



Districts for the full treatment group were pre-selected by the DOE and USAID for Punjab. Since district-level selection for the two groups was not random, equivalence at baseline of the two treatment groups cannot be assured, and a quasi-experimental design was selected. In this design, any differences in scores at baseline (and midline and endline) will be statistically removed in the analysis, i.e., the two groups will be made statistically equivalent even though their average scores may be different. This will ensure fairness in the comparison of the full and light treatment groups. In addition, scores between the groups will not be statistically tested at baseline because the goal of the evaluation is to compare the long-term progress of both groups. Providing group comparisons at baseline may introduce potential competition between the groups and invalidate the experimental design.

For the baseline assessment in Punjab, a random selection from the full treatment districts as selected for the PRP interventions resulted in the choice of Rahimyar Khan, Faisalabad, and Rawalpindi, whereas Sialkot, Jhang, and Layyah were randomly selected for the light intervention districts. For each treatment group and district, equal numbers of boys and girls schools were sampled for the EGRA testing. The sampling design met the USAID requirements of adequate sample size and equal gender representation (see the sampling section below).

Timeline

The Punjab baseline, like the other provinces, was conducted according to a timeline that started in January 2013 and ended in May 2014, with draft submissions of reports to USAID in March 2014. The final report may then be distributed to the DOE and other stakeholders as appropriate. (See Table 1 below.)

The process began in January 2013 with the planning and design of activities, including the creation of preliminary sampling designs, selection of model EGRA tasks, recruitment of staff, and budgeting/contracting. From February to August, the EGRA team, with participation from Punjab and other provinces, then prepared, piloted, and revised the EGRA tools and conducted the district/school sampling. The data collection in Punjab took place in October 2013 (and in January 2013 for Layyah district), and was followed by data entry, analysis, and reporting in March 2014. Presentations to the DOE and USAID took place in April 2014 and the final report for Punjab was completed in May 2014.

In Punjab, one district, Layyah, was tested three months after the others. While this may have slightly increased the scores in Layyah at baseline and may subsequently lower the growth estimates at midline, those small discrepancies will not invalidate the endline results. Therefore, future data collection in all of Punjab's sample districts should occur in October.

TABLE 1: TIMELINE (JANUARY 2013 TO MAY 2014)

Activity	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Plan and design EGRA activities	X	X															
Debrief to USAID and the Ministry of Education	X	X	X				X	X									
Prepare EGRA tools		X	X														
Prepare test administration manuals			X														
Train master trainers and enumerators									X								
Select and verify sample schools							X	X									
Administer EGRA										X			X				
Enter data										X	X			X			
Analyze baseline data														X			
Produce draft reports															X		
Produce presentations															X		
Disseminate draft reports																X	
Make presentations																X	
Revise and finalize reports																X	X
Submit final reports to USAID																	X

Sampling

The sampling for Punjab was finalized in July 2013 following meetings with USAID. The EGRA team conducted the school sampling in July. This included developing the sampling requirements, verifying the sample in the field, and finalizing the sample.

There were a total of 36 districts in Punjab, out of which 13 were full treatment and the remaining 23 were light treatment. The districts included in the full treatment sampled population were Bahawalpur, Dera Ghazi Khan, Faisalabad, Gujranwala, Gujrat, Kasur, Lahore, Multan, Muzaffargarh, Rahim Yar Khan, Rajanpur, Rawalpindi, and Sargodha. A simple random sample of three districts was chosen – Rahimyar Khan, Faisalabad, and Rawalpindi. The districts included in the light treatment sampled population were Bahawalnagar, Bhakkar, Chakwal, Chiniot, Hafizabad, Jhang, Jhelum, Khanewal, Khushab, Layyah, Lodhran, Mandi Bahauddin, Mianwali, Narowal, Nankana Sahib, Okara, Pakpattan, and Sahiwal. Of these Sialkot, Jhang, and Layyah were selected at random. The number of schools in the districts and the apportioned number of samples from each district by gender is shown in Table 2.

Sampling Requirements

Since the minimum requirement was 15 students per grade level in grades 3 and 5, only schools meeting that requirement were eligible for sampling. Within the treatment groups (full and light), equal numbers of boys and girls schools (35 each) were selected.

Sampling Process and Field Verification

From the chosen full treatment and light treatment districts, districts for the full and light treatment assessment groups were randomly selected. This resulted in a clustered sample. For the 35 boys and 35 girls schools in both the full and light treatment groups, the samples were divided among the selected districts according to the proportions of schools within those districts (stratified random sampling). An equal number of boys and girls schools were chosen within each group. For both groups, a second stratification was done at the "location" level, where schools were allocated by rural and urban. Table 2 shows the number of schools and replacement schools for both treatment groups per gender and location. Note that mixed schools may have been selected for some replacement schools due to not having enough options for replacement schools of strictly one gender. However, only students from the respective genders were included in those samples (i.e. if a mixed school was selected to replace a female school, only females were sampled).

TABLE 2: SCHOOLS BY DISTRICT, TREATMENT, LOCATION, AND GENDER

District	Location	Schools	Percentage	Sample Schools		Replacement Schools	
				Boys	Girls	Boys	Girls
Full Treatment Group							
Rahimyar Khan	Rural	3,293	39	14	14	4	4
Rahimyar Khan	Urban	268	3	1	1	0	0
Faisalabad	Rural	2,082	25	9	9	3	3
Faisalabad	Urban	451	5	2	2	0	0
Rawalpindi	Rural	2,048	24	8	8	3	3
Rawalpindi	Urban	299	4	1	1	0	0
Total		**8,441**	**100**	**35**	**35**	**10**	**10**
Light Treatment Group							
Layyah	Rural	1,549	26	10	10	3	3
Layyah	Urban	101	2	1	1	0	0
Jhang	Rural	1,798	29	10	10	3	3
Jhang	Urban	186	3	1	1	0	0
Sialkot	Rural	2,225	37	12	12	4	4
Sialkot	Urban	175	3	1	1	0	0
Total		**6,034**	**100**	**35**	**35**	**10**	**10**
Total (both groups)		**14,475**		**70**	**70**	**20**	**20**

Once the schools were sampled, the QCOs, supplemented by EGRA senior managers, verified the samples in the field. This step was necessary due to two factors: 1) some inaccuracies in the National Education Management Information System (NEMIS) data, and 2) changes in student numbers since the time period when the schools had submitted their data to NEMIS. If the original schools had fewer than 15 students in either grade 3 or 5, a replacement school was selected and verified. At times, schools were retained if their student numbers were near the minimum.

Intended and Actual Samples

For the full treatment group, 13 schools – five boys and eight girls – were substituted with schools randomly selected from the "replacement schools" list. The schools were replaced due to lower than expected numbers of students in the original samples. Likewise for the light treatment group, 10 schools, five boys and five girls schools, were replaced for the same reason. The actual numbers of students, teachers, and head teachers in the survey are presented in the results section.

Instrument Development

A brief summary of the instrument development process is presented below. The full results from the trans-adaptation, which involved educationists from Punjab, were presented in a report to USAID.[1]

Trans-adaptation

In February 2013, the EGRA team used tasks from recent EGRA administrations in other countries to develop a model test. Led by two international and two national assessment specialists, the EGRA team then organized a trans-adaptation workshop in Islamabad. A total of 17 English, Sindhi, and Urdu

[1] MSI (2013) *Pakistan EGRA Tools Trans-Adaptation Workshop Report*. June (Revised).

language specialists from the provincial education departments and teacher training institutes throughout Pakistan participated in the workshop. The participants were invited to 1) discuss and choose reading tasks that would be of value to the baseline assessment in Pakistan; 2) adapt each selected reading task using language appropriate content in English, Urdu, and Sindhi; and 3) ensure that the content would be suitable for grade 3 students (keeping in mind that the same tests would be used with both grade 3 and grade 5 students). The workshop resulted in pilot EGRA tests and pilot student, teacher, and head teacher questionnaires. The head teacher questionnaires included items about school characteristics.

Piloting

In March 2013, the EGRA English and Urdu tools were piloted in selected schools in AJK, ICT, and KP, while the Sindhi tools were piloted in June in Sindh. Four tools were included in the pilot: 1) a student response booklet (including the student questionnaire), 2) a student stimuli booklet, 3) a teacher questionnaire, and 4) a head teacher questionnaire. The EGRA team conducted the pilot sampling, trained the enumerators, arranged the logistics, and supervised the piloting. The team then entered the pilot data into a database, analyzed the data, and developed preliminary recommendations for final tools in preparation for the revision workshop. They also prepared a piloting report for USAID.[2]

Revision and Finalization

The EGRA team held a revision workshop in March 2013 for the Urdu and English tools with a limited number of experts from the trans-adaptation workshop. The Sindhi tools were revised in July with Sindhi language experts. Changes were made to the instruments based on the pilot data and field observations. These changes were summarized in the piloting report. The team then finalized the four instruments for each language and submitted them to USAID. USAID made suggestions, particularly around the inclusion of reading- and library-related items into the questionnaires that would provide information for the PRP and SRP. The English and Urdu instruments were approved and then used in the training workshops in advance of Round 1 data collection in May. The final instruments were comprised of the following:

- Students: 16 informational items, 8 tasks (one with 2 sub-tasks), and 34 questionnaire items.
- Teachers: 15 informational items and 52 questionnaire items.
- Head teachers: 17 informational items and 37 questionnaire items.

Data Collection

Subcontractor Selection

The EGRA team, with the participation of USAID, issued a request for proposals and followed a set of criteria to select local subcontractors for the field data collection and data entry. In August, the Society for the Advancement of Education (SAHE) was chosen for data collection activities, while the Basic Education for Awareness, Reforms and Empowerment (BEFARe) was selected for data entry activities. MSI, STS, and SAHE collaborated on the data collection in Punjab.

Data Collection

In September, EGRA senior managers trained MTs and QCOs during a two-week session in Islamabad. The MTs then spent one week in Islamabad training the SAHE data collection team, which consisted of one regional coordinator, six field supervisors, and 64 enumerators. The QCOs, coordinators, supervisors, and enumerators organized the logistics for the data collection. Following the training and logistical preparations, the QCOs and field supervisors conducted a three-day refresher course for the enumerators in each district just prior to commencing data collection in the schools.

[2] MSI (2013). *Pakistan EGRA Instrument Development and Pilot Data Analysis*.

Due to a sampling issue during which a full treatment district, Bahawalpur, was chosen as part of the light treatment group, additional data collection was organized in January 2014 for a new light treatment district, Layyah, which was randomly selected. Over a 12-day period in October (and a six-day period in January for data collection in the Layyah district), enumerators spent a day in each of the 140 schools to collect the baseline data in Punjab. The enumerators were in regular communication with the EGRA senior manager, QCOs, coordinator, and field supervisors to check on the status of data collection and to troubleshoot any issues. After collecting the data from the schools, the enumerators submitted their booklets to the supervisors and QCOs for verification and feedback. At the end of data collection, all booklets were returned to Islamabad for data entry.

Data Entry

Data Entry

In May, the EGRA team developed a customized data entry application so 1) the exact data from the booklets and questionnaires could be entered into a database, and 2) the computers used for data entry could be networked with a server. In September, the team trained the BEFARe data coordinator, four supervisors, and 36 data entry operators. In October and November, the EGRA and BEFARe teams did the data entry for over 23,000 student booklets, along with the questionnaires for the students, teachers, and head teachers (Rounds 2 and 3). This included roughly 4,200 booklets for Punjab, and an additional 660 booklets, approximately, for Layyah district in February 2014.

Data Cleaning

In November 2013 and February 2014, the EGRA and BEFARe teams conducted the data verification and reconciliation. Following USAID requirements, 100 percent of the data were entered twice (double data entry) and any discrepancies between the first and second databases were reconciled. A clean data file was then provided to the data analysis team.

Data Analysis

Methodology

In June, the EGRA statisticians and psychometrician from STS developed a research plan that included the following steps: 1) reliability estimates, 2) task and item statistics, 3) mean and grand mean scores (percent correct scores), 4) data plots, 5) timed and untimed task scores, and 6) questionnaire results. They used SPSS for the analysis. Some of the analyses were replicated to ensure that the calculations were accurate. Descriptive analyses and inferential statistical comparisons were conducted by grade level and gender for the student scores, and for the three sets of questionnaire data.

Validity and Reliability

Validity evidence for the tests was derived from previous experiences with EGRA in other developing countries, as well as through the trans-adaptation process in Pakistan. The test developers targeted grade 3 for the level of the tasks. The assumption was that the grade 5 students should perform better than the grade 3 students on each of the tasks.

For reliability, a generally accepted method is to estimate the internal consistency reliability (Coefficient Alpha) of the test. The minimum reliability threshold is approximately 0.75 to 0.80 for tests of this nature. Reliability was estimated for each province. Table 3 shows the reliability estimates in Punjab for grades 3 and 5 were 0.85 and 0.84, respectively. These reliabilities are excellent and lend credibility for the tests' internal consistency, indicating that the items are generally measuring similar reading constructs for both tests.

TABLE 3: RELIABILITY ESTIMATES

Language	Grade Level	Tasks	N-count	Alpha
Urdu	Grade 3	9	2,079	0.85
	Grade 5	9	2,069	0.84

Note that there were actually eight tasks, but one of the tasks (Task 7) was administered and scored in two parts, so the equivalent of nine tasks were used for the analysis.

Score Calculation

The EGRA data were analyzed three ways. First, p-values and item-total correlations were generated for assessing the difficulty and quality of the items and tasks. Second, the percent correct for each task provided an indication of the Punjab students' mastery of the tasks; since this metric was the same across all tasks, the percent correct scores for each of the tasks were combined to produce a reading summary score (i.e., a total score). Third, timed task scores were used to calculate Punjab students' fluency. Each of these analyses is explained below.

Item P-values and Item-Total Correlations

P-values and item-total correlations are classical test theory statistics that are used to evaluate the performance of individual items and the tasks they comprise. Item difficulty is measured by p-values, which range from 0.00 to 1.00. Higher p-values indicate easier items because a higher percent of students posted correct responses. The other classical statistic is the item-total correlation, and it ranges from -1.00 to +1.00. This statistic measures how close the item or task relates to the overall percent correct on the summary score. Values above 0.20 are an indication of a good quality item or task.

Percent Correct

The results of the EGRA testing were calculated using task and summary scores. Table 4 lists the tasks, stimuli, raw score ranges, and the method for calculating the task and summary scores on the test. For each of the tasks, the stimuli (items) (i.e., questions, letters, sounds, words, and non-words) were worth one score point. The score points were added and since the range of raw scores varies across the tasks, the percent of correct scores was used to report all results. No weighting was used with the tasks to calculate the summary scores. Each task summary score was calculated using the total of the percent correct scores and dividing it by the number of items (i.e., the average percent correct). The overall Reading Summary Score was calculated by adding all of the task summary scores and dividing by nine (total number of tasks) to arrive at the average.

Timed Tasks Scores

The scores on the timed tasks were calculated (adjusted) by taking the number of correct responses (i.e., the raw score) times 60 seconds then dividing that number by the number of seconds used to read the stimulus. For instance, if a student read 75 letters correctly in 30 seconds, their letters-correct-per-minute score would be 150 (75 words x 60 seconds/30 seconds). Given another example, if a student read 50 words correctly in 30 seconds, his or her timed task score would be 100 words per minute (50 words x 60 seconds/30 seconds). Table 4 lists the number of stimuli per task. Recall the percent correct scores ranged from zero to 100. The method for calculating phonics and fluency scores yielded much higher maximum values, upwards of 200 at baseline (see task box plots in Annex 2, Figures A1-A2).

TABLE 4: EGRA SCORE RANGES AND CALCULATIONS

Task (Subtest)	Stimuli	Score Range	Calculation
1. Orientation to print	5 questions	0-5	Percent correct of answers
2. Letter name recognition	100 letters (timed)	0-100	Percent correct of letters
3. Phonemic awareness	10 questions	0-10	Percent correct of words
4. Letter sound knowledge	100 sounds (timed)	0-100	Percent correct of sounds
5. Familiar word reading	50 words (timed)	0-50	Percent correct of words
6. Non-word reading	50 non-words (timed)	0-50	Percent correct of non-words
7a. Passage reading	60 words (timed)	0-60	Percent correct of words
7b. Passage comprehension	5 questions	0-5	Percent correct of answers
8. Listening comprehension	3 questions	0-3	Percent correct of answers
Reading Summary Score	9 tasks	0-100	Average of percent correct

An example of percent correct scores for each of the tasks and as a summary score is provided below. The raw score is divided by the maximum score (the highest score possible in the score range) to produce the percent correct score for each task. Then, the task scores are averaged to produce the summary score. Note that each of the task percent correct scores is weighted equally to provide the summary score.

TABLE 5: EXAMPLE OF EGRA PERCENT CORRECT AND SUMMARY SCORES

Task (Subtest)	Maximum Score	Raw Score	% Correct Score
1. Orientation to print	5	3	60.0%
2. Letter name recognition	100	68	68.0%
3. Phonemic awareness	10	5	50.0%
4. Letter sound knowledge	100	42	42.0%
5. Familiar word reading	50	34	68.0%
6. Non-word reading	50	25	50.0%
7a. Passage reading	60	50	83.3%
7b. Passage comprehension	5	2	40.0%
8. Listening comprehension	3	1	33.3%
Reading Summary Score	--	--	55.0%

An example of timed task scores (adjusted) is provided below for the five fluency tasks. The formula explained above is used (timed task score = raw score x 60 seconds/seconds used).

TABLE 6: EXAMPLE OF EGRA TIMED TASK SCORES

Task (Subtest)	Raw Score	Seconds Used	Timed Task Score
2. Letter name recognition	68	48	85.0
4. Letter sound knowledge	42	60	42.0
5. Familiar word reading	34	48	42.5
6. Non-word reading	25	40	37.5
7a. Passage reading	50	40	75.0

CHAPTER 3: FINDINGS AND RESULTS

This chapter presents the findings and results from the EGRA baseline in Punjab. There are sections on the student sample, task and item statistics, score calculation, task and summary scores, timed task scores, and questionnaire findings.

Student Sample

The intended sample was 70 full and 70 light treatment schools. Within these schools, the target was to assess 15 students in each grade, totaling 4,200 students (i.e., 2,100 for each gender, each treatment group, and each grade). Table 7 shows the number of students in the sample by grade and gender. The column labeled "Missing" is the number of children who did not fill in the gender question.

For the full treatment group in grades 3 and 5, the actual samples were 98.7 and 98.5 percent of the intended sample, respectively. For the light treatment group the actual sample size was 98.6 and 98.7percent for grades 3 and 5, respectively. The entire grade 3 sample was 98.7 percent, and grade 5 was 98.1 percent. The boys' percent (99.6) was slightly higher than the girls' (97.3). A small number of students in grade 3 (n = 7) and grade 5 (n = 5) did not complete the gender item on the questionnaire. Due to the missing gender codes, when analyzing the students by this characteristic the sample was 4,136 students, 98.5 percent of the intended 4,200 sample records. However, when the data were not analyzed by gender, the total actual sample was higher by 12 students at 4,148 pupils (98.8 percent of the intended 4,200 students).

TABLE 7: ACTUAL STUDENT SAMPLE BY GRADE AND GENDER

Treatment	Grade Level	Sample	Boys	Girls	Missing	Total
Full Treatment	Grade 3	Students	517	519	6	1,042
		% of Target	98.5%	98.9%	--	99.0%
	Grade 5	Students	516	518	2	1,036
		% of Target	98.3%	98.7%	--	98.7%
	Total	Students	1,033	1,037	8	2,078
		% of Target	98.4%	98.8%	--	99.0%
Light Treatment	Grade 3	Students	539	497	1	1,037
		% of Target	102.7%	94.7%	--	98.8%
	Grade 5	Students	520	510	3	1,033
		% of Target	99.0%	97.1%	--	98.4%
	Total	Students	1,059	1,007	4	2,070
		% of Target	100.9%	95.9%	--	98.6%
Full and Light Treatment	Grade 3	Students	1,056	1,016	7	2,079
		% of Target	100.6%	96.8%	--	99.0%
	Grade 5	Students	1,036	1,028	5	2,069
		% of Target	98.7%	97.9%	--	98.5%
	Total	Students	2,092	2,044	12	4,148
		% of Target	99.6%	97.3%	--	98.8%

Task and Item Statistics

Table 8 shows the statistics for the tasks for the Punjab sample. Two classical statistics are provided: p-values and item-total correlations. P-values indicate the average score of the students on the tasks, or the difficulty of the tasks for the students. The item-total correlations in the table are actually task-total correlations, which indicate the degree to which the tasks can discriminate between low- and high-achieving students; this is an indicator of the quality of the items. P-values can range from 0.00 to 1.00, with higher values indicating easier items. Item-total correlations can range from -1.00 to +1.00, with values above +0.20 indicating that the item (or task) is of good quality.

In Table 8 below, the task p-values for grade 3 in Punjab ranged from 0.11 to 0.45, thus providing a spread on the lower half of the difficulty spectrum. The p-values for grade 5 were higher, ranging from 0.20 to 0.76 or in the middle parts of the range. The level of difficulty for both grade levels was appropriate for this baseline measure because there will be enough room in the scale for capturing growth during the midline and endline assessments. For item-total correlations, a generally acceptable threshold is 0.20 and above. All of the task scores in grades 3 and 5 had item-total correlations greater than 0.25, indicating very good quality for these tasks. Complete item statistics are provided in Annex 1 at the end of this report.

TABLE 8: TASKS STATISTICS (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3		Grade 5	
	P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print	0.41	0.25	0.53	0.35
2. Letter name recognition	0.43	0.62	0.55	0.53
3. Phonemic awareness	0.33	0.28	0.44	0.31
4. Letter sound knowledge	0.11	0.36	0.20	0.28
5. Familiar word reading	0.42	0.84	0.76	0.80
6. Non-word reading	0.23	0.81	0.50	0.76
7a. Passage reading	0.45	0.84	0.76	0.78
7b. Passage comprehension	0.16	0.75	0.44	0.72
8. Listening comprehension	0.28	0.44	0.46	0.41

Task and Summary Scores

The next part of the analysis involved plotting the summary scores. Histograms of the summary scores (Figures 2 and 3) show that the distributions are moving from left to right from grade 3 to grade 5, which is strong evidence that the children are learning basic skills at the primary school level. In addition, as with the task and item statistics, it also shows that there is room for growth at each grade level. The main goal of the intervention is to see movement of the score distributions to the right within the same grade level (i.e., grades 3 and 5) from the baseline to midline to endline.

FIGURE 2: GRADE 3 SUMMARY SCORES

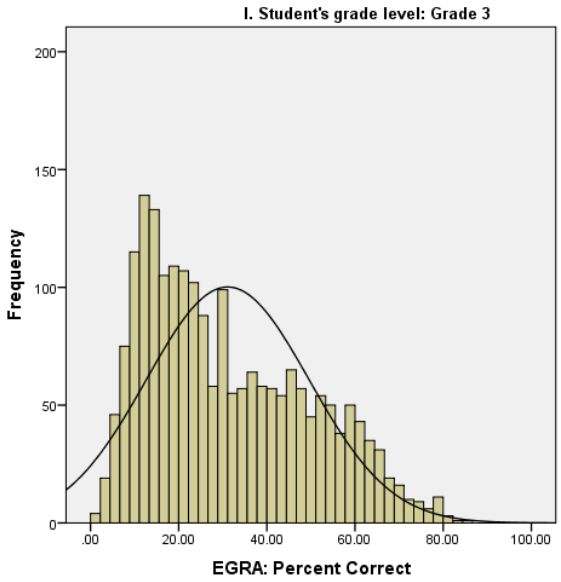


FIGURE 3: GRADE 5 SUMMARY SCORES

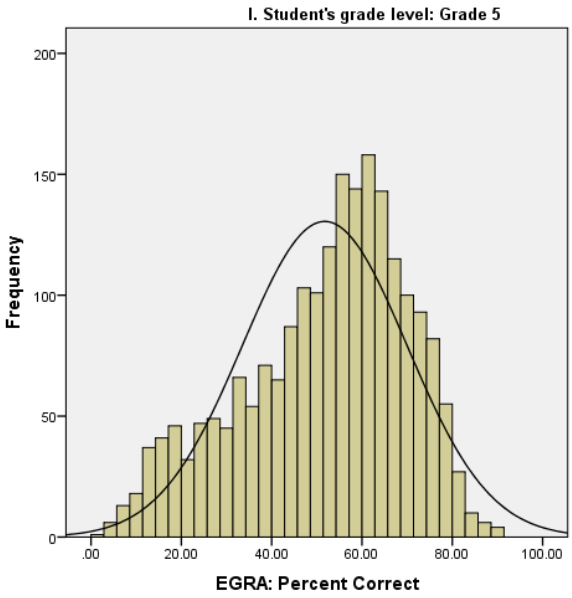


Table 9 and 10 and Figure 4 provide the average scores by task using percent correct scores. The score for each task was calculated using the total number correct and dividing by the number of items. For instance, a student who scored 3 out of 5 on Task 1 would receive a score of 60 percent. Averages were then calculated for all students on Task 1, which in Punjab was 40.7 percent for grade 3 and 53.0 percent for grade 5. The same type of calculation was made for each student and each task. The table also includes the differences from grade 3 to grade 5, e.g., 53.0 percent minus 40.7 percent equals 12.3 percentage points.

Grade 3 posted the highest scores in passage reading, familiar word reading, letter name recognition, and orientation to print, though the percent correct scores were all below 50 percent. The most difficult tasks for these students were comprehension (passage and listening) and phonics (non-word reading, letter sound knowledge, and phonemic awareness). The older students' scores showed a similar pattern. The highest grade 5 scores were in familiar word reading and passage reading; whereas the most challenging tasks were comprehension (passage and listening) and letter sound knowledge. The most challenging reading tasks for the Punjab students were reading comprehension and phonics, particularly letter sound knowledge.

There was also substantial progression from grade 3 to grade 5 on the summary score (17.6 points). The phonics tasks of letter sound knowledge (9.0 points) and phonemic awareness (11.5 points) showed the least improvement. In contrast, the greatest gains were in familiar word reading (34.5 points), passage reading (31.5 points), passage comprehension (28.3 points), and non-word reading (27.0 points). In areas where there are large differences, interventions at grade 3, or earlier, could have particularly large effects in accelerating children's learning. These gains were also consistent for both treatment groups (Table 10).

TABLE 9: PERCENT CORRECT SCORES BY GRADE AND TASK (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3	Grade 5	Difference (G5 – G3)
1. Orientation to print	40.7%	53.0%	12.3% points
2. Letter name recognition	42.7%	55.0%	12.3% points
3. Phonemic awareness	33.0%	44.5%	11.5% points
4. Letter sound knowledge	10.6%	19.6%	9.0% points
5. Familiar word reading	41.5%	76.0%	34.5% points
6. Non-word reading	23.1%	50.1%	27.0% points
7a. Passage reading	44.5%	76.0%	31.5% points
7b. Passage comprehension	15.5%	43.8%	28.3% points
8. Listening comprehension	28.4%	46.0%	17.6% points
Reading Summary Score	31.1%	51.7%	20.6% points

The full treatment group had higher scores in all nine tasks for both grades (Table 10), with the exception of grade 3 in familiar word reading. The total differences by group were small (2.2 points at each grade level). This slight discrepancy will be corrected statistically at midline and endline by analyzing the growth for each group from baseline to midline and endline. Because this is a baseline report, the group differences were not statistically tested at this time.

TABLE 10: PERCENT CORRECT SCORES BY GRADE, TASK, AND GROUP

Task (Subtest)	Full		Light	
	Grade 3	Grade 5	Grade 3	Grade 5
1. Orientation to print	42.7%	53.8%	38.7%	52.2%
2. Letter name recognition	44.8%	58.1%	40.4%	52.0%
3. Phonemic awareness	34.3%	46.7%	31.8%	42.3%
4. Letter sound knowledge	8.1%	14.3%	12.9%	24.9%
5. Familiar word reading	41.6%	76.9%	41.5%	75.1%
6. Non-word reading	24.2%	52.6%	22.0%	47.6%
7a. Passage reading	45.4%	78.5%	43.5%	76.7%
7b. Passage comprehension	16.5%	45.6%	14.4%	42.9%
8. Listening comprehension	31.6%	48.8%	25.1%	43.1%
Reading Summary Score	32.2%	52.8%	30.0%	50.6%

FIGURE 4: FULL TREATMENT PERCENT CORRECT SCORES BY GRADE AND TASK

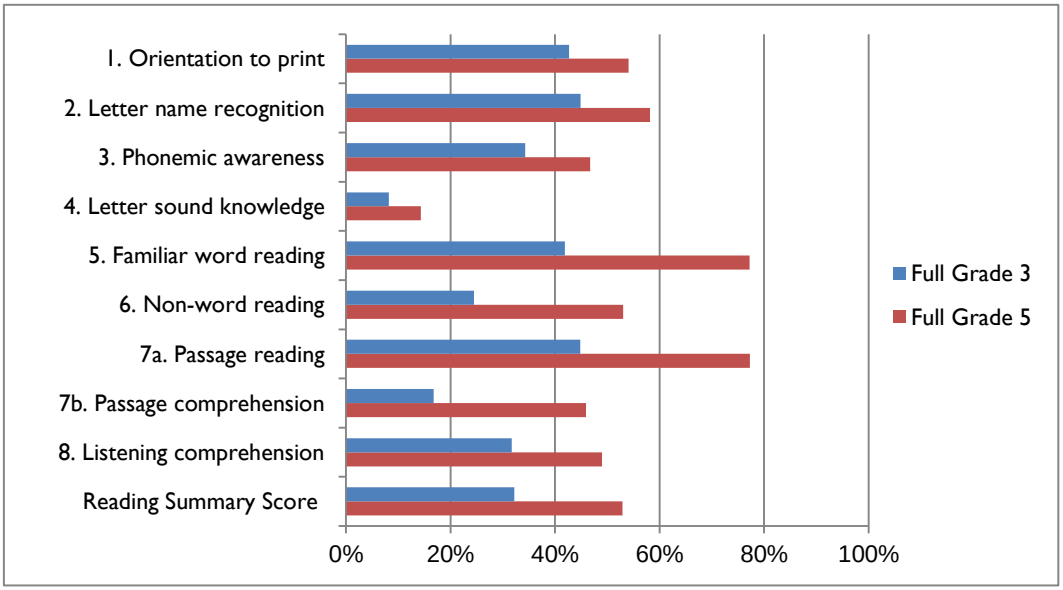
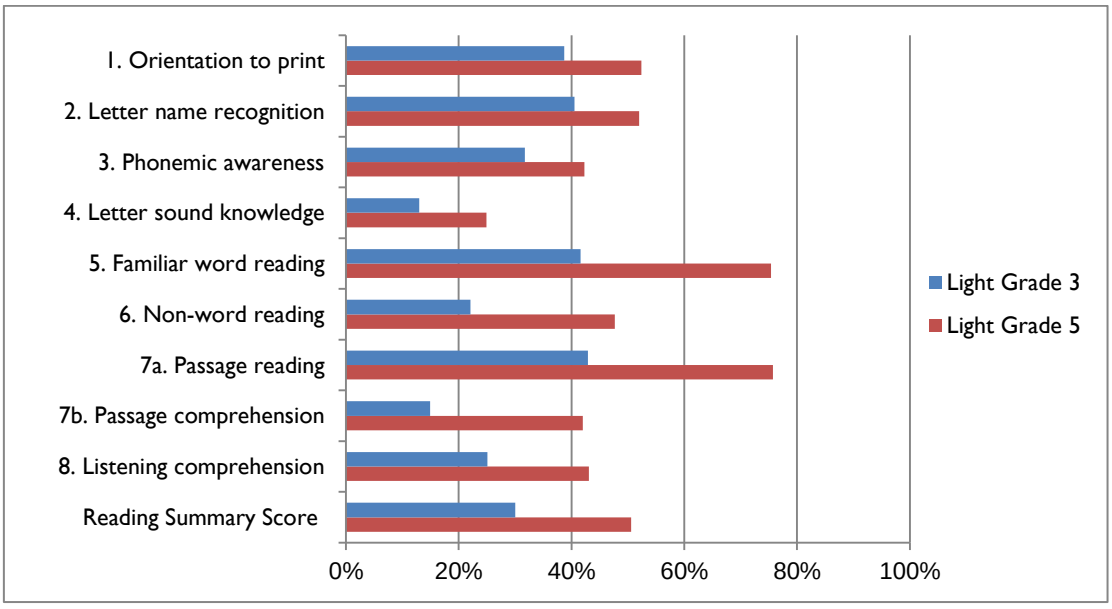


FIGURE 5: LIGHT TREATMENT PERCENT CORRECT SCORES BY GRADE AND TASK



Percent correct scores by grade and gender are presented in Table 11 and in Figures 6 and 7. Boys and girls showed different patterns in reading skills. At grade 3, boys performed best on orientation to print (42 percent), letter name recognition (37.6 percent), and passage reading (35.7 percent). Grade 3 girls displayed higher scores in passage reading (53.7 percent) and familiar word reading (50.2 percent), followed by letter name recognition (48.0 percent) and orientation to print (39.5 percent). At grade 5, boys and girls were best at familiar word reading (69.9 percent and 82.3 percent, respectively) and passage reading (71.4 percent and 84.1 percent, respectively). Boys had difficulty with passage comprehension (33.0 percent) and letter sound knowledge (18 percent), while girls were mostly challenged by letter sound knowledge (21.2 percent). In comparing scores between the genders, girls' scores were significantly higher ($p < 0.01$) on the EGRA summary score on all tasks with the exception of orientation to print at grade 5

(higher, but not significantly so). Again, the data show that students in Punjab have the most difficulty with phonics, especially letter sound knowledge, and comprehension.

TABLE 11: PERCENT CORRECT SCORES BY GRADE, TASK, AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
1. Orientation to print	42.0%	39.5%	52.4%	54.1%
2. Letter name recognition	37.6%	48.0%*	51.1%	59.2%*
3. Phonemic awareness	30.5%	35.6%*	41.0%	47.9%*
4. Letter sound knowledge	8.9%	12.4%*	18.0%	21.2%*
5. Familiar word reading	33.1%	50.2%*	69.9%	82.3%*
6. Non-word reading	18.7%	27.8%*	44.3%	56.1%*
7a. Passage reading	35.7%	53.7%*	71.4%	84.1%*
7b. Passage comprehension	10.6%	21.0%*	33.0%	55.0%*
8. Listening comprehension	26.3%	35.3%*	43.8%	48.4%*
Reading Summary Score	27.0%	35.4%*	47.2%	56.4%*

*Indicates that the performance of the group was significantly higher, p< 0.01

FIGURE 6: GRADE 3 PERCENT CORRECT SCORES BY TASK AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

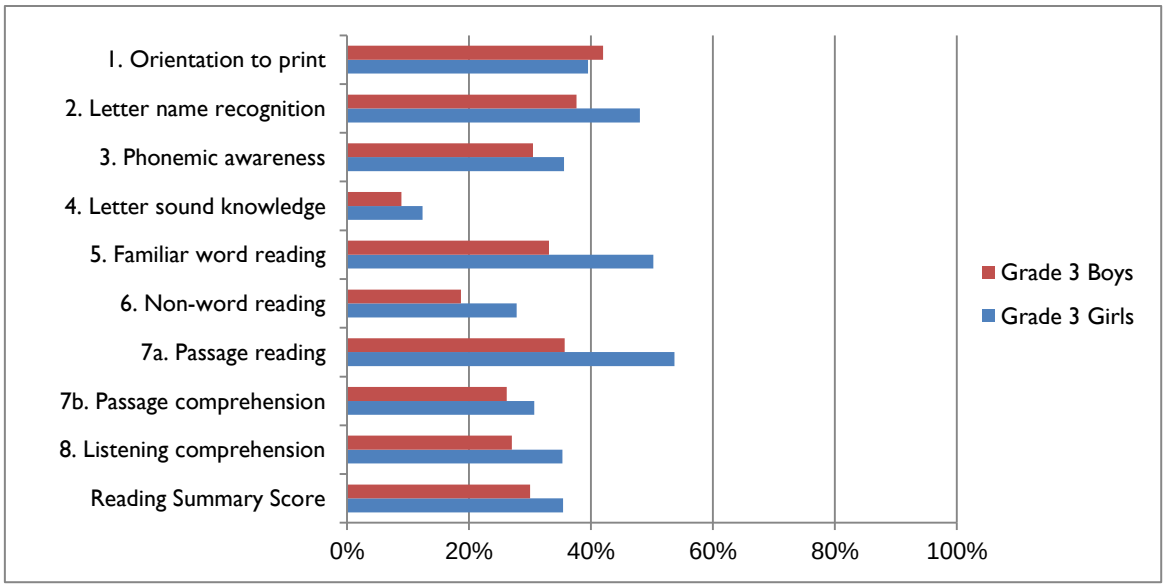
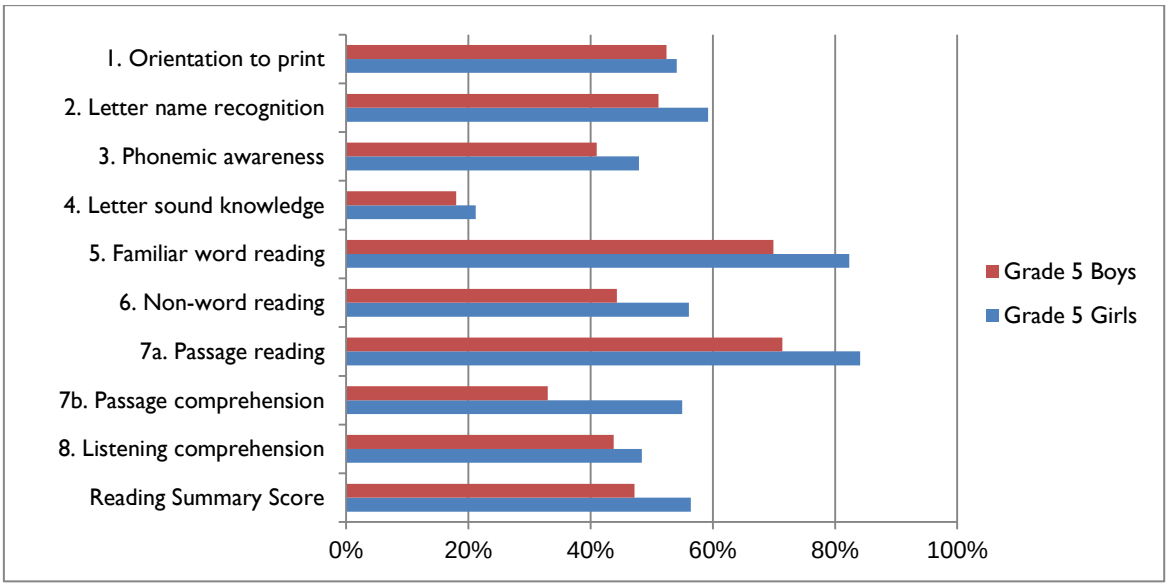


FIGURE 7: GRADE 5 PERCENT CORRECT SCORES BY TASK AND GENDER (FULL AND LIGHT TREATMENT GROUPS)



The final table in this section (Table 12) further disaggregates the scores by treatment group, grade level, and gender. As seen in the tables above, the full treatment group scored higher on most of the tasks, which will be statistically corrected at the midline and endline. By gender, the girls in both full and light treatment groups generally scored higher than the boys.

TABLE 12: PERCENT CORRECT SCORES BY GROUP, GRADE, AND GENDER

Task (Subtest)	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
1. Orientation to print	44.3%	41.4%	52.3%	55.9%	39.9%	37.5%	52.6%	52.4%
2. Letter name recognition	39.3%	50.4%	52.9%	63.4%	35.9%	45.4%	49.2%	54.9%
3. Phonemic awareness	32.1%	36.5%	42.7%	50.7%	29.0%	34.7%	39.4%	45.1%
4. Letter sound knowledge	7.3%	9.2%	12.7%	16.0%	10.5%	15.7%	23.3%	26.5%
5. Familiar word reading	33.1%	50.7%	69.2%	85.3%	33.3%	50.3%	71.3%	79.5%
6. Non-word reading	19.3%	29.6%	44.4%	61.7%	18.2%	26.1%	44.5%	50.7%
7a. Passage reading	35.8%	53.8%	69.8%	84.8%	34.5%	52.0%	70.7%	80.7%
7b. Passage comprehension	11.2%	22.3%	33.8%	58.1%	9.9%	19.6%	32.2%	51.9%
8. Listening comprehension	28.3%	35.1%	46.0%	52.2%	24.2%	26.1%	41.6%	44.5%
Reading Summary Score	27.9%	36.6%	47.1%	58.7%	26.2%	34.2%	47.2%	54.0%

Timed Tasks: Phonics and Reading-Rate Fluency Scores

Fluency is a measure of reading efficiency. On the Pakistan EGRA there were two types of fluency measures, phonics and reading rate. The phonics-fluency subtest included letter name recognition, letter

sound knowledge, and non-word reading, whereas, the reading-rate fluency subtest consisted of familiar word and passage reading.

Tables 13 to 17 below show scores in terms of raw scores (instead of the percent correct scores on the previous tables). Table 13 has the maximum raw scores attained by students on each task at each grade level. Tables 14 to 17 have mean scores for the students. In addition, adjustments were made to the raw scores for those students who finished the task before the end of one minute. For instance, if a student read 50 words correctly in 30 seconds, their words correct per minute score would be 100 (50 words x 60 seconds/30 seconds). Because these calculations are different from percent correct, the maximum scores are higher (see Figures A1-A2 in Annex 2). Table 13 provides the baseline maximum scores at grades 3 and 5 for the five timed tasks.

Table 13 displays the maximum score achieved in Punjab for each fluency task. These figures are much higher that the percent correct calculations because they are based on the number of words successfully read in a minute. These maximum scores should provide a reference for comparing the mean scores listed in Tables 14 to 16. Please note that maximum scores can contain extremely high scores, called outliers. Item 6 in Table 13 shows that the grade 3 scores were higher than grade 5, but that is due to one student in grade 3 with a very high score of 131; the second highest score was 93.

TABLE 13: BASELINE MAXIMUM SCORES ON FLUENCY (TIMED) TASKS (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3	Grade 5
2. Letter name recognition	125	127
4. Letter sound knowledge	100	128
6. Non-word reading	131	108
Reading-Rate Fluency Subtest	**Grade 3**	**Grade 5**
5. Familiar word reading	143	166
7a. Passage reading	177	180

As shown in Table 14, grade 3 students had the best success at naming letters. They were less successful at identifying letter sounds and reading non-words. For both grades, the phonics task of non-word reading was especially difficult. Students in grade 5 performed better on the reading-rate fluency tasks, and like grade 3 students, had difficulty with the phonics tasks, especially identifying sounds and non-word reading. The greatest gains from grade 3 to 5 were in the reading-rate fluency tasks. In contrast, the lack of growth in phonics fluency should be a target for instruction.

TABLE 14: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
2. Letter name recognition	43.1	55.7	12.6 points
4. Letter sound knowledge	23.7	34.8	11.1 points
6. Non-word reading	21.8	36.3	14.5 points
Reading-Rate Fluency Subtest	**Grade 3**	**Grade 5**	**Difference (G5 – G3)**
5. Familiar word reading	37.9	64.0	26.1 points
7a. Passage reading	35.0	81.7	46.7 points

In comparing the treatment groups (Table 15), there were only small differences in the fluency tasks between the two groups. Both groups showed similar patterns in fluency. Letter sound knowledge and non-word reading were the most challenging, while letter name recognition and reading-rate fluency tasks were less demanding. Again, these minor differences will be corrected statistically at the midline and endline evaluations.

TABLE 15: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GROUP

Phonics Fluency Subtest	Grade 3		Grade 5	
	Full	Light	Full	Light
2. Letter name recognition	45.3	40.9	59.0	52.4
4. Letter sound knowledge	22.9	24.3	31.9	36.8
6. Non-word reading	21.9	21.6	37.3	35.2
Reading-Rate Fluency Subtest	**Grade 3**		**Grade 5**	
	Full	**Light**	**Full**	**Light**
5. Familiar word reading	37.1	38.7	64.9	63.1
7a. Passage reading	34.9	35.1	81.9	81.6

Girls' fluency rates were higher than the boys on all tasks for both grades (Table 16). All comparisons were statistically significant ($p < 0.01$). The largest differences were in passage reading at both grades. This table also highlights that both genders had difficulty with the phonics tasks of letter sound knowledge and non-word reading.

TABLE 16: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	38.1	48.1*	51.6	59.8*
4. Letter sound knowledge	20.5	26.9*	32.3	37.3*
6. Non-word reading	19.2	24.1*	32.4	39.9*
Reading-Rate Fluency Subtest	**Grade 3**		**Grade 5**	
	Boys	**Girls**	**Boys**	**Girls**
5. Familiar word reading	32.7	42.2*	57.5	70.2*
7a. Passage reading	26.0	44.4*	69.5	94.1*

*Indicates that the performance of the group was significantly higher, $p< 0.01$

The final table in this section (Table 17) further disaggregates the scores by treatment group, grade level, and gender. As with the percent correct scores, the full treatment group scored higher on many of the tasks, which will be statistically corrected at the midline and endline.

TABLE 17: PHONICS AND READING-RATE FLUENCY TASKS MEANS BY GROUP, GRADE, AND GENDER

Phonics Fluency Subtest	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
2. Letter name recognition	39.9	50.5	53.5	64.4	36.5	45.6	49.8	55.1
4. Letter sound knowledge	19.9	25.9	29.6	34.0	20.8	27.5	34.0	39.8
6. Non-word reading	19.3	24.3	32.5	41.4	19.1	23.8	32.2	38.2
Reading-Rate Fluency Subtest	**Full Treatment**				**Light Treatment**			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
5. Familiar word reading	31.9	41.2	56.6	72.8	33.4	43.3	58.5	67.6
7a. Passage reading	25.7	44.1	67.0	96.9	26.1	44.7	72.1	91.3

Questionnaire Findings

Selected results are presented below, including for those characteristics or items that showed significant differences in student scores. Due to the students having the same language, the results were combined for the full and light treatment groups to increase the sample size and more accurately detect effects between the categories. Note that there were some students, teachers, and head teachers who did not respond to certain questionnaire items; they were labeled as missing. The total averages for the summary scores were calculated based on those who responded.

Since these are baseline data, reporting on the full and light treatment groups together will not affect the analyses at midline and endline. We combined the survey data for the groups since some of the questions led to reporting by relatively small categories (e.g., for teacher qualifications) and we wanted to know whether the survey results were associated with the student scores in general.

In addition, since the samples were by treatment group, the results will be generalized to the populations for each group. This will be done prior to the midline. The results will be generalized to by calculating sampling weights, applying the weights to the results, and then generalizing to the population by treatment group. We will also do this for the midline and endline. The current analyses only apply to the sampled districts.

Statistical significance was determined based on *t*-tests for indicators with two categories and analyses of variance for indicators with three or more categories (with post-hoc pairwise comparisons). The significance value was set at $p < 0.05$; a 95 percent confidence level, but if a higher significance level was found (e.g., $p < 0.01$), then the more significant value is listed.

Student Questionnaires

One survey question asked the students what language was spoken in the home. Both grades showed similar patterns in the primary language spoken at home (Table 18). Most families spoke Punjabi (51 percent), followed by Other (25 percent), Urdu (21 percent), and Pashto (1 percent). Less than 1 percent of the students spoke English, Sindhi, and Balochi at home. Although the assessments were in Urdu, about 21 percent of third grade and 17 percent of fifth grade Punjab students spoke Urdu as their primary language in the home.

TABLE 18: PERCENTAGE OF STUDENTS BY LANGUAGE SPOKEN AT HOME

Language	Grade 3		Grade 5	
	n-count	Percent	n-count	Percent
English	17	0.8%	2	0.1%
Urdu	439	21.1%	348	16.8%
Sindhi	7	0.3%	3	0.1%
Pashto	26	1.3%	32	1.5%
Punjabi	1,068	51.4%	1,172	56.6%
Balochi	7	0.3%	8	0.4%
Other or Missing	515	24.8%	524	24.4%
Total	2,079	100.0%	2,069	100.0%

Table 19 lists summary scores by student age. According to the National Education Policy (2009), the official age of the students at the beginning of the different grade levels of primary education is 6 to 10 years old. Since the baseline took place during the school year, the normal ages for this analysis were set at 8 to 9 years old for grade 3 and 10 to 11 years old for grade 5. The students were placed into three categories: younger than normal age for their grade, normal age, and older than normal age. At grade 3, there were no significant differences among the age groups. Conversely, at grade 5, the older-than-normal scores were significantly lower than the normal-age scores ($p < 0.05$). Note that this finding is contrary to what was found in the other provinces, where there were significant differences at grade 3 (older students with higher scores), but no differences at grade 5.

TABLE 19: SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	149	28.6%	144	53.1%
Normal age	1,040	31.6%	1,011	52.7%
Older than normal age	886	30.9%	911	50.4%*
Missing	4	--	3	--
Total	2,079	31.1%	2,069	51.7%

* Indicates that the performance of the group was significantly lower, $p < 0.05$ level

Table 20 shows the summary scores according to whether the student reads the Quran at home. There were significant differences at grade 5 favoring students who read the Quran.

TABLE 20: SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	64	27.2%	57	44.9%
Yes	1,981	31.3%	1,995	52.0%*
Missing	34	--	17	--
Total	2,079	31.1%	2,069	51.7%

* Indicates that the performance of the group was significantly higher, $p < 0.01$

Table 21 depicts the differences in scores based on whether there is a library at the school. Students reporting the presence of a library did not have significantly higher scores in either grade. Please note that about 5 percent of students had missing responses or did not know if the school had a library.

TABLE 21: SUMMARY SCORES BY THE PRESENCE OF A LIBRARY AT THE SCHOOL

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	680	31.4%	658	51.8%
Yes	1,284	31.1%	1,311	51.9%
Missing	115	--	100	--
Total	2,079	31.1%	2,069	51.7%

In Tables 22 to 24, the data showed that the existence of newspapers and magazines generally made a difference in reading scores for both grades. Surprisingly, having books in the house was not related to higher reading scores.

TABLE 22: SUMMARY SCORES BY THE PRESENCE OF NEWSPAPERS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,603	30.1%	1,398	50.7%
Yes	476	34.3%*	671	53.9%*
Missing	0	-	0	-
Total	2,079	31.1%	2,069	51.7%

* Indicates that the performance of the group was significantly higher, p< 0.01

TABLE 23: SUMMARY SCORES BY THE PRESENCE OF MAGAZINES AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,855	30.6%	1,745	51.3%
Yes	224	35.2%*	324	54.3%*
Missing	0	--	0	--
Total	2,079	31.1%	2,069	51.7%

* Indicates that the performance of the group was significantly higher, p< 0.01

TABLE 24: SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,034	32.0%	1,173	51.6%
Yes	1,045	30.2%	896	51.9%
Missing	0	--	0	--
Total	2,079	31.1%	2,069	51.7%

* Indicates that the performance of the group was significantly higher, p< 0.05

The final set of student questions (in Tables 25 to 27) pertained to children's reading habits at home. In general, unlike other provinces, having someone read to children at home, having children read to someone else at home, and children reading silently at home was not related to higher reading scores. The only exception was grade 5 students who read silently at home and had significantly ($p < 0.05$) higher scores than their peers who did not read silently at home.

TABLE 25: SUMMARY SCORES BY CHILDREN HAVING SOMEONE READ TO THEM AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	808	30.7%	773	51.6%
Yes	1,265	31.4%	1,284	51.9%
Missing	6	--	12	--
Total	2,079	31.1%	2,069	51.7%

* Indicates that the performance of the group was significantly higher, p< 0.05

TABLE 26: SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	770	30.6%	701	51.3%
Yes	1,297	31.5%	1,364	51.9%
Missing	12	--	4	--
Total	2,079	31.1%	2,069	51.7%

* Indicates that the performance of the group was significantly higher, p< 0.01

TABLE 27: SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	496	31.0%	503	50.0%
Yes	1,569	31.2%	1,561	52.4%*
Missing	14	--	5	--
Total	2,079	31.1%	2,069	51.7%

*Indicates that the performance of the group was significantly higher, p< 0.05

Teacher Questionnaires

With the smaller sample size, the analysis of the teacher questionnaires was limited to descriptive statistics, i.e., no group comparisons. There were 116 teachers who taught grade 3 and 111 who taught grade 5. Responses from some teachers who indicated that they taught both grades are not included since all results are presented by grade.

Tables 28 and 29 provide information on teacher academic and professional qualifications, neither of which showed consistent patterns in the student scores.

TABLE 28: SUMMARY SCORES BY TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc./M.Phil.	28	31.5%	39	52.1%
B.A./B.Sc.	27	33.4%	30	53.0%
F.A./F.Sc.	17	31.0%	19	50.6%
Matric	30	31.2%	22	52.3%
Missing	1	--	1	--
Total	116	31.1%	111	51.7%

TABLE 29: SUMMARY SCORES BY TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	13	33.1%	18	51.3%
B.Ed.	34	33.5%	33	53.7%
C.T.	14	27.2%	9	49.2%
P.T.C.	52	32.4%	47	52.4%
Missing	3	--	3	--
Total	116	31.1%	111	51.7%

In an analysis of student scores by teacher age and experience, there were no consistent patterns of younger or older teachers, or those with less or more experience, relating to lower or higher student scores (Tables 30 and 31). Again, small teacher sample sizes made drawing conclusions difficult.

TABLE 30: SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	43	30.7%	35	49.9%
Between 41 and 50	51	32.2%	59	53.7%
51 and more	22	33.3%	17	52.2%
Missing	0	--	0	--
Total	116	31.1%	111	51.7%

TABLE 31: SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	30	30.2%	27	49.5%
Between 11 and 20	29	33.6%	32	53.4%
Between 21 and 30	42	32.0%	43	54.3%
31 or more	9	33.0%	7	45.5%
Missing	6	--	2	--
Total	116	31.1%	111	51.7%

There were no significant differences in summary scores at grade 3 or grade 5 among teachers who attended or did not attend in-service training sessions (Table 32). Once more, any differences should be interpreted with caution due to the small sample size.

TABLE 32: SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	52	32.3%	47	53.8%
One time	33	30.4%	36	52.8%
Two times	12	33.5%	15	46.5%
Three times	10	34.1%	9	52.2%
Missing	9	--	4	--
Total	116	31.1%	111	51.7%

Head Teacher Questionnaires

Similar to the teachers, the sample size for the head teacher questionnaires was small, so data interpretations should be treated with caution. Tables 33 and 34 show average reading scores by the academic background and professional qualification of head teachers. In terms of academic and professional qualification, no discernible pattern relationships were found with summary reading scores.

TABLE 33: SUMMARY SCORES BY HEAD TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc./M.Phil.	86	31.7%	86	52.5%
B.A./B.Sc.	29	33.6%	29	52.2%
F.A./F.Sc.	10	25.4%	10	47.7%
Matric	15	26.5%	15	49.5%
Missing	0	--	0	--
Total	140	31.1%	140	51.7%

TABLE 34: SUMMARY SCORES BY HEAD TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	54	33.9%	54	53.8%
B.Ed.	44	29.6%	44	50.8%
C.T.	13	33.0%	13	52.0%
P.T.C.	27	27.5%	27	49.4%
Missing	2	--	2	--
Total	140	31.1%	140	51.7%

Tables 35 and 36 provide information on head teachers' experience and in-service training. For both grades, no evident pattern was revealed in the head teachers' years of experience or in-service training. Again, any differences should be interpreted with caution due to the small sample size.

TABLE 35: SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	45	31.8%	45	51.6%
3 to 5	19	27.8%	19	50.3%
6 to 10	16	28.3%	16	46.7%
11 or more	48	32.5%	48	53.7%
Missing	12	--	12	--
Total	140	31.1%	140	51.7%

TABLE 36: SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	57	32.0%	57	52.2%
1 time	62	30.9%	62	51.8%
2 times	12	32.0%	12	52.4%
More than 2 times	8	27.9%	8	49.3%
Missing	1	--	1	--
Total	140	31.1%	140	51.7%

Tables 37 and 38 provide data on head teachers' support to teachers in reading and the training that head teachers received in teaching reading. Summary scores were not significantly higher for head teachers who reported supporting teachers in reading or who attended reading training sessions.

TABLE 37: SUMMARY SCORES BY HEAD TEACHER SUPPORT OF TEACHERS IN READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	39	31.6%	39	50.2%
Yes	101	31.0%	101	52.3%
Missing	0	--	0	--
Total	140	31.1%	140	51.7%

TABLE 38: SUMMARY SCORES BY HEAD TEACHER TRAINING IN TEACHING READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	63	31.1%	63	50.4%
Yes	74	31.4%	74	53.1%
Missing	3	-	3	--
Total	140	31.1%	140	51.7%

School Characteristics

The final section provides information on school characteristics (from the head teacher questionnaires) by student summary scores. As with the teacher and head teacher characteristics, most patterns appeared to be inconclusive (Tables 39 to 43). The few urban schools (8) in the sample had higher reading scores than those in rural settings (62). Again, this finding should be interpreted with caution due to the low sample size of urban schools. Girls and mixed-gender schools performed better at grade 3, while the boys and mixed-gender schools posted higher scores at grade 5. Over 90 percent of the schools reported having a PTA and approximately 50 percent stated they had a school library. Neither of these factors, however, was related to reading scores. Lastly, the 87 percent of schools who reported having better infrastructure also had higher student reading scores; scores increased with the addition of water, electricity, or toilets.

TABLE 39: SUMMARY SCORES BY SCHOOL LOCATION

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Rural school	62	31.0%	62	52.4%
Urban school	8	41.6%	8	56.7%
Missing	0	--	0	--
Total	70	31.1%	70	51.7%

TABLE 40: SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Boys school	32	26.9%	32	58.6%
Girls school	29	35.9%	29	46.1%
Mixed Gender	9	38.9%	9	58.2%
Missing	0	--	0	--
Total	70	31.1%	70	51.7%

TABLE 41: SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	5	41.6%	5	56.3%
Yes	64	31.4%	64	52.3%
Missing	1	--	2	--
Total	70	31.1%	70	51.7%

TABLE 42: SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	22	33.7%	22	51.2%
Yes	48	31.5%	48	53.6%
Missing	0	--	0	--
Total	70	31.1%	70	51.7%

TABLE 43: SUMMARY SCORES BY INFRASTRUCTURE (DRINKING WATER, ELECTRICITY, TOILETS)

Number of Infrastructures (Water, Electricity, Toilets)	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	0	--	0	--
1	1	--	1	--
2	8	29.1%	8	52.6%
3	61	32.2%	61	52.7%
Missing	0	--	0	--
Total	140	31.1%	140	51.7%

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

This final chapter provides conclusions and recommendations from the Punjab EGRA baseline. The conclusions are organized according to the two main sections in the report: 1) design and methodology, and 2) findings and results. There are also recommendations based on the instrument development, data collection, data entry, and analysis.

Design and Methodology

1. The design followed USAID evaluation guidelines for a cross-sectional approach. This will allow for an examination of the progress of students in grades 3 and 5 over the life of the PRP. Though the language of instruction in Punjab was recorded as English in the NEMIS, the decision was made to conduct the assessment in Urdu since the transition to English-medium has not yet been made in schools in Punjab. In addition, Punjab has two treatment groups: full and light. This will allow for an evaluation of the full treatment effects above and beyond those of the light treatment.

2. The sampling issues were addressed as well as could have been expected. In a limited number of schools, there was an issue of a lack of the requisite number of students per grade level. The actual sample of schools was 100 percent and the actual sample of students reached 99 percent of the intended sample.

3. The EGRA test in Urdu administered in Punjab was of good quality. The reliability estimates were in the high part of the range (alpha = 0.85, and 0.84, for grades 3 and 5 respectively) of previous EGRA administrations in other countries. The task statistics were acceptable, with an appropriate range of p-values and item-total correlations that were at an acceptable level of quality. The characteristics of the tests were such that it should be a strong measure of potential progress over time due to project-led interventions. Given the baseline scores, histograms, and box plots, the EGRA is expected to accurately measure the higher reading abilities that are expected at midline and endline.

4. The field implementation was successful, though there were difficulties to overcome, including the low actual enrollment of students in some schools and the sampling issue during the selection of the light treatment districts. There was a high level of standardization reported by the quality control officers, which they attributed to the effective training process by the EGRA team. The team paid careful attention to detail in the logistics and test administration, which was reflected in the low error rates in the booklets and in the data entry.

Findings and Results

The Punjab evaluation involves two kinds of analyses: 1) a comparison of full and light treatment groups to determine the effects of full treatment above and beyond that of the light treatment, and 2) a comparison of each group to itself at the baseline, midline, and endline. Please see Figure 1 and the accompanying text for a fuller description of the evaluation design.

Several key findings emerged from the baseline assessment in Punjab. These are as follows:

1. EGRA was administered to a robust sample of 2,079 grade 3 students and 2,069 grade 5 students. The test reliability was excellent for both grades (alpha = 0.85 for grade 3 and alpha = 0.84 for grade 5). These high reliabilities indicate that the items worked well in measuring reading constructs at both grades.

2. The task and item statistics showed that EGRA discriminates well between low- and high-achieving students in both grades. The task p-values for grade 3 provided a spread on the lower to lower-middle section of the difficulty range, while p-values for grade 5 were higher and covered the upper-lower half to the high-middle parts of the spectrum. All task scores at grades 3

and 5 had item-total correlations equal to or greater than 0.25, indicating good discrimination quality for these tasks. (Complete item statistics are listed in Annex 1.)

3. Grade 3 students posted the highest scores in passage reading, familiar word reading, letter name recognition, and orientation to print, however, the percent correct scores were all below 50 percent. The most difficult tasks for the students in Punjab were comprehension (passage and listening) and phonics (non-word reading, letter sound knowledge, and phonemic awareness). The grade 5 students' scores showed similar patterns. At grade 5, the highest scores were in familiar word reading and passage reading; whereas the most challenging tasks were comprehension (passage and listening) and letter sound knowledge. The most challenging reading tasks for the students in Punjab were reading comprehension and phonics, particularly letter sound knowledge.

4. There was substantial progression from grade 3 to grade 5 on the summary score (17.6 points). The phonics tasks of letter sound knowledge (9.0 points) and phonemic awareness (11.5 points) showed the lowest improvement. In contrast, the greatest gains were in familiar word reading (34.5 points), passage reading (31.5 points), passage comprehension (28.3 points), and non-word reading (27.0 points). In areas where there are large differences, interventions at grade 3, or earlier, could have particularly large effects in accelerating children's learning. This progress was consistent across gender and treatment groups.

5. Of the nine task percent correct scores, the full treatment group had higher scores on eight tasks for both grades, but these differences were small. This slight discrepancy will be corrected statistically at midline and endline by analyzing the growth for each group from baseline to midline and endline. Because this is a baseline report, the group differences were not statistically tested at this time.

6. Boys and girls showed different patterns in reading skills. At grade 3, boys performed best on orientation to print (42.0 percent), letter name recognition (37.6 percent), and passage reading (35.7 percent). Grade 3 girls displayed higher scores in familiar word reading (50.2 percent) and passage reading (53.7 percent), followed by letter name recognition (48.0 percent), and orientation to print (39.5 percent). At grade 5, boys and girls were best at familiar word reading (69.9 percent) and passage reading (71.4 percent), but boys had difficulty with passage comprehension (33.0 percent) and letter sound knowledge (18 percent), while girls were mostly challenged by letter sound knowledge (21.2 percent). In comparing scores between the genders, the girls' scores were significantly higher ($p < 0.001$) on the EGRA summary score on all tasks except orientation to print at grade 5 (higher, but not significantly). Again, the data show that students in Punjab have the most difficulty with letter sound knowledge and comprehension.

7. Students were timed on five tasks as they read letters, words, or passages. These tasks were categorized into phonics (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Students at both grades had lower phonics scores than reading-rate fluency scores. Moreover, gains from grade 3 to grade 5 were lower for phonics than for reading-rate fluency tasks. Girls' fluency rates were significantly higher than the boys for all tasks ($p < 0.001$). There were only small differences between the light and full treatment groups' fluency scores. Although the passage was designed for grade 3, this difference shows that the reading-rate fluency levels in grade 3 are low, but that students can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn. Specifically mastery of phonics and phonemic awareness should help the students become better overall readers. It is clear that this type of knowledge and these skills are not receiving an appropriate emphasis in schools in Punjab.

8. From the student questionnaires, the results showed that, unlike other provinces, having someone read to children at home, having children read to someone else at home, and children reading silently at home were not related to higher reading scores. The only exception was grade 5 students who read silently at home had significantly ($p < 0.05$) higher scores than their peers who did not read silently at home.

9. School, teacher, and head teacher questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of variation in the scores that were related to their characteristics. For example, an analysis of student scores by teacher and head teacher education, certification, age, experience, and attendance at in-service trainings found no consistent patterns relating to lower or higher student scores.

10. For the schools, the few urban schools (8) had higher reading scores than those in rural settings (62), however, this finding should be interpreted with caution due to the low number of urban schools in the sample. Girls and mixed-gender schools performed better at grade 3, while the boys and mixed-gender schools posted higher scores at grade 5. Over 90 percent of the schools reported having a PTA and approximately 50 percent stated they had a school library. Neither of these factors, however, was related to higher reading scores. Lastly, the 87 percent of schools who reported having better infrastructure also had higher student reading scores; scores increased with the addition of water, electricity, or toilets.

Evaluation Recommendations

Given the success of the baseline assessment in Punjab (and in the other provinces), the methods used in 2013 should be repeated as much as possible for the midline and endline assessments in future years.

This should be conducted as follows:

1. The EGRA instruments proved to be of high quality, and equivalent versions of those tools should be developed – through trans-adaptation, piloting, and revision – for the midline and endline assessments so that progress can be accurately measured over time.

2. The EGRA items and tasks had good discrimination (quality) values and covered the low-to-middle part of the difficulty range. At baseline, the reading scores were relatively low for both grades and show room for growth. In addition, histograms and box plots provided evidence that the tool is expected to measure higher levels of reading-rate fluency that are anticipated following project-led interventions. Therefore, the baseline data indicates that EGRA is appropriate for measuring increases in reading ability at midline and endline.

3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. The data collection procedures were very effective, as a high percentage of the targeted students were tested. They should be maintained in the midline and endline, i.e., keep the same districts and schools, along with the sampling methods at the school level.

4. Because of an issue during sampling, one of the light treatment districts, Layyah, was tested three months after the others. While this may have slightly increased the scores in Layyah at baseline and subsequently may decrease the gains at midline, those small discrepancies will not invalidate the endline results. Therefore, pending discussions with the USAID Lahore team, we recommend that future data collection in all of Punjab's sample districts should occur in October.

5. The systems for field data collection should be replicated, with the same systems for recruitment and training for the master trainers, field supervisors, quality control officers, and enumerators as used in the baseline.

6. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.

7. The analysis should follow the same procedures, with calculations of reliability, difficulty, task percent-correct scores, summary scores, and timed task scores. The baseline, midline, and endline scores should be computed using the same procedures so that improvements in students' reading can be accurately examined over time.

8. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. The baseline EGRA data can be used for establishing these reading proficiency levels.

9. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires to collect data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

In general, the Punjab baseline was successful in providing accurate data on which to base decisions for implementation of the PRP interventions, and also for tracking student reading progress over time. It provides a solid foundation for the midline and endline assessments.

ANNEXES

Annexes 1 to 4 provide additional information on the EGRA baseline. Specifically, the annexes have the following:

Annex 1 gives complete item statistics – p-values (the difficulty of the items) and item-total correlations (the quality of the items) by grade – for the items associated with the various tasks. These are more detailed than the task statistics presented in Chapter 3 of the report. Measurement specialists often request these kinds of item statistics for the purposes of quality control, analysis, and test equating.

Annex 2 provides box plots for the fluency tasks. The box plots are more task-specific than the overall score distributions (histograms) presented in the report. They show the median (middle score), the range (highest and lowest scores), and the distribution of scores (by quartiles) for each task. The task-specific distributions are useful to EGRA specialists who place emphasis on the fluency tasks.

Annex 3 gives two examples of categorizing passage reading fluency scores using performance levels. The categorizations – along with raw scores and scale scores -- are often used to interpret test scores. The first example combines reading speed with comprehension, while the second example only uses reading speed. Each example uses a set of cut-scores for placing the students into performance categories.

Annex 4 provides detailed information on the second example, with results for each category of fluency and each level of comprehension. These data can be used as evidence on the reliability of using a combined measure of fluency and comprehension for setting performance cut-scores. The validity of combining these scores is more of an issue for reading experts.

Annex 1: Complete Item Statistics by Grade

Table A1 presents item statistics for the untimed tasks, each of which have multiple items. For instance, task 1 (orientation to print) has item statistics for its five items (Q1 to Q5). Note that the timed tasks are lists of letters, sounds, and words, i.e., not items, so it is not necessary to calculate item statistics for them.

Previously, we presented task statistics (Chapter 3, Table 8) with explanations of how they are calculated. These item statistics are calculated in the same way. They show the difficulty and quality of the items. Recall that when constructing a test, we strive for tasks and items that have difficulty values (p-values) that are spread across the range from about 0.05 to 0.90 and quality values (item-total correlations) of at least 0.20. The difficulty values ranged from 0.05 to 0.67 for grade 3 and 0.08 to 0.80 for grade 5, indicating a strong range of item difficulties. A total of 20 and 19 items for grades 3 and 5 respectively out of the 23 items per grade had item-total correlations of at least 0.20, indicating high quality items.

TABLE A1: COMPLETE ITEM STATISTICS BY GRADE

Task (Subtest)	Item	Grade 3		Grade 5	
		P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	Q1	0.67	0.25	0.71	0.28
	Q2	0.44	0.27	0.57	0.28
	Q3	0.34	0.18	0.40	0.18
	Q4	0.11	0.05	0.30	0.12
	Q5	0.48	0.08	0.69	0.19
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.51	0.34	0.67	0.43
	Q2	0.29	0.36	0.47	0.52
	Q3	0.31	0.28	0.39	0.33
	Q4	0.25	0.32	0.39	0.45
	Q5	0.34	0.28	0.43	0.41
	Q6	0.41	0.29	0.55	0.42
	Q7	0.24	0.29	0.30	0.38
	Q8	0.26	0.33	0.37	0.41
	Q9	0.23	0.32	0.31	0.40
	Q10	0.47	0.33	0.57	0.39
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.15	0.56	0.45	0.49
	Q2	0.09	0.44	0.30	0.42
	Q3	0.06	0.41	0.18	0.35
	Q4	0.26	0.59	0.65	0.56
	Q5	0.21	0.61	0.63	0.58
8. Listening comprehension (untimed)	Q1	0.27	0.29	0.50	0.24
	Q2	0.05	0.22	0.08	0.15
	Q3	0.54	0.25	0.80	0.24

Annex 2: Box Plots for Phonics and Reading-rate Fluency Tasks

EGRA places a high emphasis on fluency (timed) tasks. In addition to the descriptive statistics in Table 9 (percent correct scores) and Table 14 (fluency task means), we show box plots for the different fluency tasks. Widely used since their development in the 1960s, box plots are a convenient way for graphically presenting numerical data.

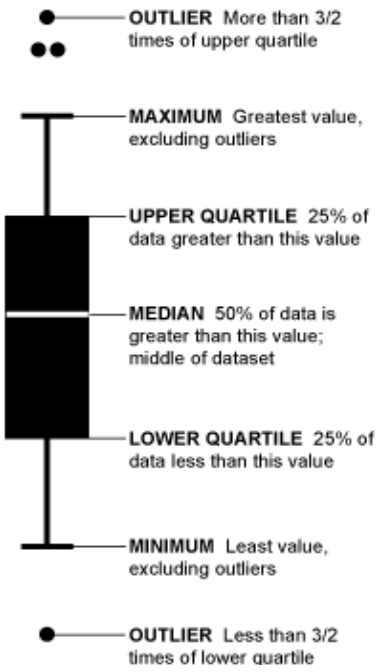
Box plots have two characteristics: 1) central tendency (i.e., the median, or the middle score in the data) and 2) variation (i.e., the range, with scores grouped by quartile). The boxes (which are actually rectangles) represent the two middle quartiles of the scores and the "whiskers" represent the upper and lower quartiles. The small circles on the ends of the whiskers represent outliers. The figure below provides a more detailed explanation for interpreting box plots.

FIGURE A1: UNDERSTANDING BOXPLOTS

Median
The median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.

Upper quartile
Seventy-five percent of the scores fall below the upper quartile.

Lower quartile
Twenty-five percent of scores fall below the lower quartile.



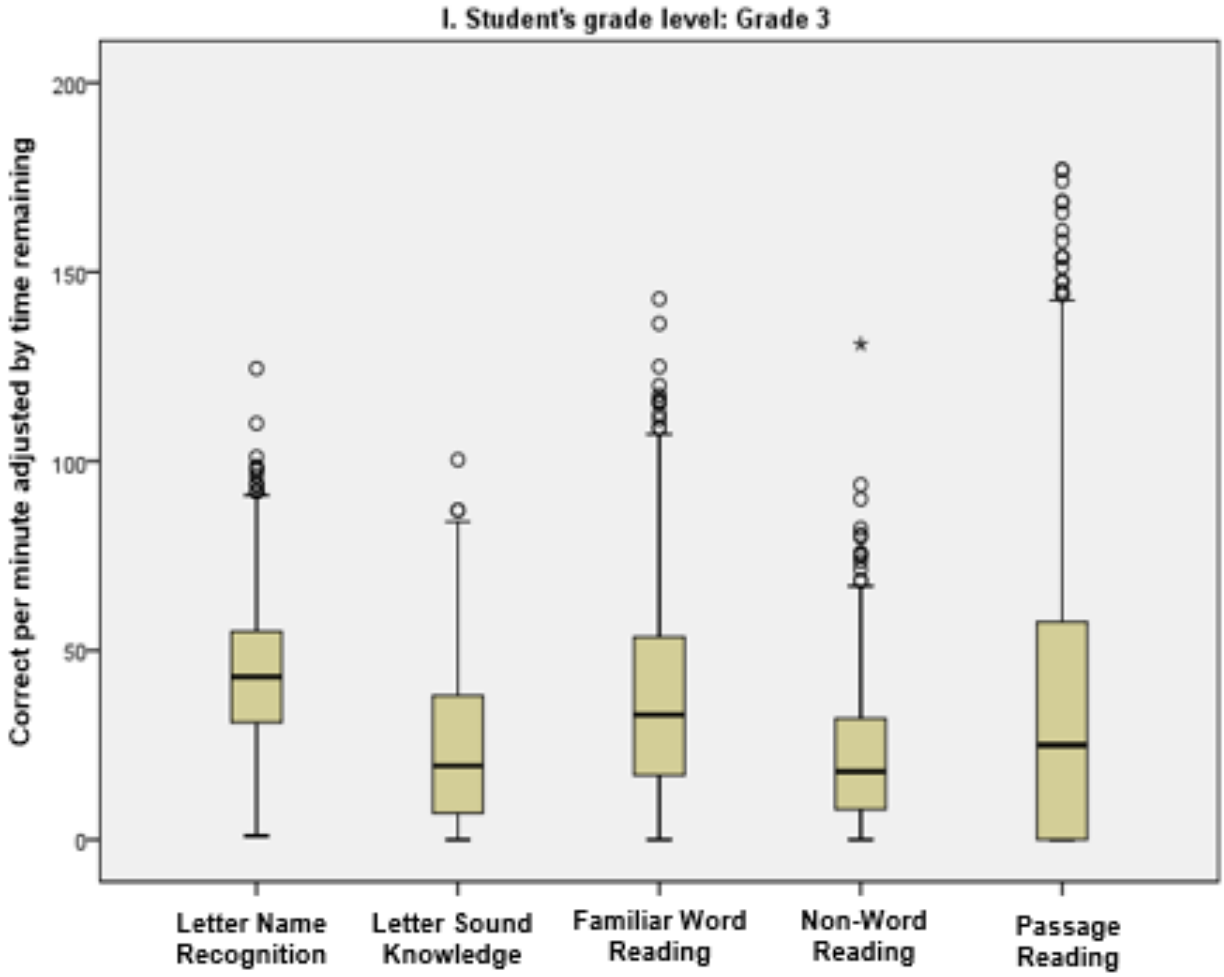
Box plots are presented below (Figures A2 and A3) for the results by grade level on the five fluency (timed) tasks: letter name recognition (task 2), letter sound knowledge (task 4), familiar word reading (task 5), non-word reading (task 6), and passage reading (task 7a).

Grade 3

For grade 3, the central tendency (i.e., the median speed, or the line in the middle) for each of the tasks ranged from about 20 (letter sound knowledge) to about 40 (letter name recognition) items per minute. It shows that the students had much better knowledge of letter names than letter sounds.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 80 (letter sound knowledge) to about 150 (passage reading). It shows that the scores were more spread out when reading words (familiar word reading and passage reading) than letters (letter name recognition and letter sound knowledge).

FIGURE A2: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 3



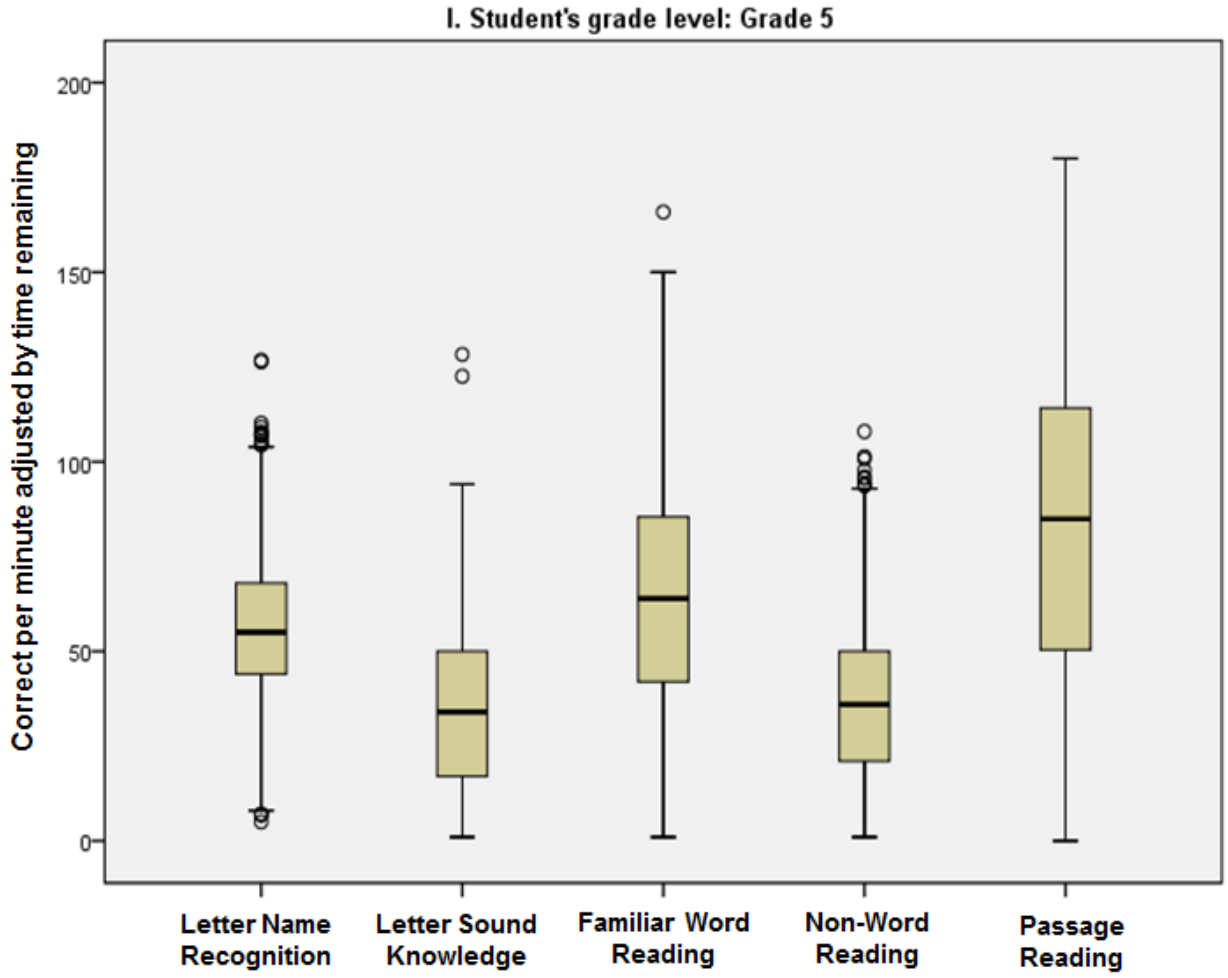
Grade 5

For grade 5, the central tendency (the median speed) for each of the tasks ranged from about 30 (letter sound knowledge) to about 80 (passage reading) items per minute. It shows that the students had better knowledge of words (familiar word reading and passage reading) than sounds (letter sound knowledge and non-word reading).

The variation (range of scores) for each of the tasks varied from about 130 (letter sound knowledge) to about 180 (passage reading). It shows that the scores were more spread out when reading words (familiar word reading and passage reading) than sounds (letter sound knowledge and non-word reading).

Note also that the medians and the ranges increased from grade 3 to grade 5 for all fluency tasks. Many students are becoming more fluent readers at grade 5, but there are also those students who are either non-readers or very low readers. These children lack of knowledge of letter names, sight words, connected text, and (especially) phonics.

FIGURE A3: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 5



Annex 3: Examples of Fluency Score Threshold Calculations

There are different ways of interpreting test scores. Three of the main ways are 1) raw scores (e.g., number correct), 2) scale scores (e.g., percent correct), and 3) percentile scores (e.g., rank in relation to other students). In the report, we presented scores in terms of number correct (for the fluency tasks) and percent correct (for all tasks). We could also calculate the percentile scores for each student, though this is not normally done with EGRA. Note that these kinds of calculations do not change or affect the actual results, but they do involve issues of interpretability.

A fourth main way of interpreting scores is through performance categories, e.g., low, middle, and high. This requires setting cut-scores, or thresholds, to separate the student scores into categories, e.g., two cut-scores lead to three performance categories. The following analysis shows two examples of calculating thresholds for passage reading scores (WCPM), which allows us to place the student scores into different performance categories. Note that performance categories are often accompanied by performance level descriptors (PLDs), which give a text-based explanation of the meaning of the scores in each category. We have not developed PLDs for these examples since 1) the threshold setting is at a preliminary stage and 2) reading specialists with knowledge of local curricula and context generally develop the PLDs.

Fluency using an 80 percent comprehension threshold

In the first example, we used a method that has been suggested by some EGRA specialists. It involves calculating the mean reading speed associated with 80 percent comprehension and then applying it as a fluent cut-score. In other words, the mean reading speed for these students signifies whether the students are fluent readers through using both passage reading speed *and* comprehension in the calculation; the fluent cut-score separates the fluent readers from the non-fluent readers. To establish a second threshold, we again followed the suggested method and used the lowest level of reading (1 WCPM) as the non-fluent cut-score. The two cut-scores resulted in three performance levels: non-readers (low), non-fluent readers (middle), and fluent readers (high).

At grade 3, the mean reading speed on the passage reading task for students who scored 80 percent on the passage comprehension task was 96 WCPM. With this method, 96 WCPM becomes a threshold for grade 3 students who are fluent using both passage reading speed *and* comprehension. At grade 5, the mean score on the passage reading task for students who scored 80 percent on the passage comprehension task was 114 WCPM. Then 114 WCPM becomes a threshold for grades 5 students who are fluent using both passage reading speed *and* comprehension.

The definitions of the three categories in terms of WCPM and the percentages of grades 3 and 5 students in the categories are shown in Table A2 below.

TABLE A2: THRESHOLDS FOR WCPM WITH 80 PERCENT COMPREHENSION

Category (Performance Level)	Grade 3		Grade 5	
	WCPM	% of Students	WCPM	% of Students
Non-Reader	0	26.5%	0	6.9%
Non-Fluent Reader	1 to 95	69.4%	1 to 113	66.8%
Fluent Reader	96 and above	4.0%	114 and above	26.3%
Total	--	100.0%[1]	--	100.0%

([1]) Due to rounding, the total sums to 99.9%

Note that the majority (over two-thirds) of the students are in the middle category at each grade level. This is due the large range of scores for this category, i.e., from the students who score just above non-readers to those who score just below fluent readers are in the non-fluent reader (middle) category.

Fluency using fixed interval thresholds

In the second example, we used fixed intervals of WCPM for the performance levels. This reduced the problem of having a large range of students in the middle category by creating early reader and intermediate reader categories. It also follows common practice when setting performance categories of having between three and five levels for student scores. We used an interval of 40 WCPM to produce five performance levels, along with a category for the non-readers. The five levels were: non-readers (0 WCPM); early readers (1-40 WCPM); intermediate readers (41-80 WCPM); fluent readers (81-120 WCPM); and advanced readers (121 and above WCPM).

TABLE A3: THRESHOLDS FOR WCPM WITH FIXED INTERVALS

Category (Performance Level)	WCPM	% of Students	
		Grade 3	Grade 5
Non-Reader	0	26.5%	6.9%
Early Reader	1 to 40	35.7%	13.3%
Intermediate Reader	41 to 80	25.7%	25.7%
Fluent Reader	81 to 120	9.4%	33.1%
Advanced Reader	121 and above	2.7%	21.0%
Total	--	100.0%	100.0%

At both grades 3 and 5, the fixed interval method allowed for more distribution of the scores across the categories. We can also see a shift in percentages of students in each category from grade 3 to grade 5; the performance categories allow for a score interpretation showing that students are improving across the grade levels, with more scores in the lower categories at grade 3 and more scores in the higher categories at grade 5.

Remarks

While it is possible to use such percentages to set cut-scores for interpretation purposes at the baseline, midline and endline, this analysis should be taken as preliminary. For instance, more well-known and accepted method of setting thresholds – which is commonly called "standard setting" by measurement specialists – involve holding a workshop with local reading experts to set the cut-scores according to the experts' conceptions of what students should know and be able to do in order to be classified into a performance category. There are several well-known methods, e.g., Angoff and Bookmark, which have been judged as valid and reliable for this purpose.[3] Further discussions on setting thresholds involving local reading experts are recommended.

[3] References include: Zieky, M. & Perie, M. (2006). *A primer on setting cut-scores on tests of educational achievement*. Princeton, New Jersey: Educational Testing Service; Cizek, G. (1996). *Standard-setting guidelines*. Educational Measurement: Issues and Practices, Spring 1996, p. 13-21; Cizek, G., Bunch, M., & Koons, H. (2004). *Setting performance standards: Contemporary methods*. Educational Measurement: Issues and Practices, Winter 2004.

Annex 4: Distribution of Reading Fluency and Comprehension Scores using Fixed Intervals

In this last annex, we provide more information on the relationship between reading fluency (speed) and comprehension using information from the fixed interval method. While the data show a positive relationship between speed and comprehension, there are sizeable numbers of "fluent" readers with little comprehension. Our conclusion is that setting a cut-score using a less than reliable indicator, such as the mean speed of students with 80 percent comprehension (i.e., using *both* speed and comprehension), can be problematic. The result is categorizing some students as fluent readers who in fact, according to the definition, are not, i.e., they have high reading speed but low comprehension. It may be better to set thresholds based solely on a single indicator – reading speed – rather than mixing it with comprehension.

The figures and tables below (Tables A4-A5 and Figures A4-A5) expand on the data in Table A3. They show the results for reading fluency (in terms of speed) by comprehension level for grades 3 and 5. We used the categories based on intervals of 40 WCPM, along with a category for the non-readers (0 WCPM). Comprehension levels were calculated in terms of percent correct scores (e.g., 20 percent is the same as correctly answering one question out of five total questions). For instance, at grade 3, 100 percent of the non-readers have 0 percent comprehension and 14 percent of the advanced readers have 100 percent percent comprehension. Also, at grade 3, 23 percent (15 percent + 8 percent) of the fluent readers and 48 percent (34 percent + 14 percent) of the advanced readers have comprehension levels of 80 percent or above.

TABLE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

Category (Performance Level)	WCPM	% of Students by Comprehension Level						
		0%	20%	40%	60%	80%	100%	Total
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	84%	13%	3%	1%	0%	0%	100%
Intermediate Reader	41 to 80	29%	28%	22%	16%	5%	2%	100%
Fluent Reader	81 to 120	5%	14%	25%	34%	15%	8%	100%
Advanced Reader	121 and above	7%	4%	16%	25%	34%	14%	100%

FIGURE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

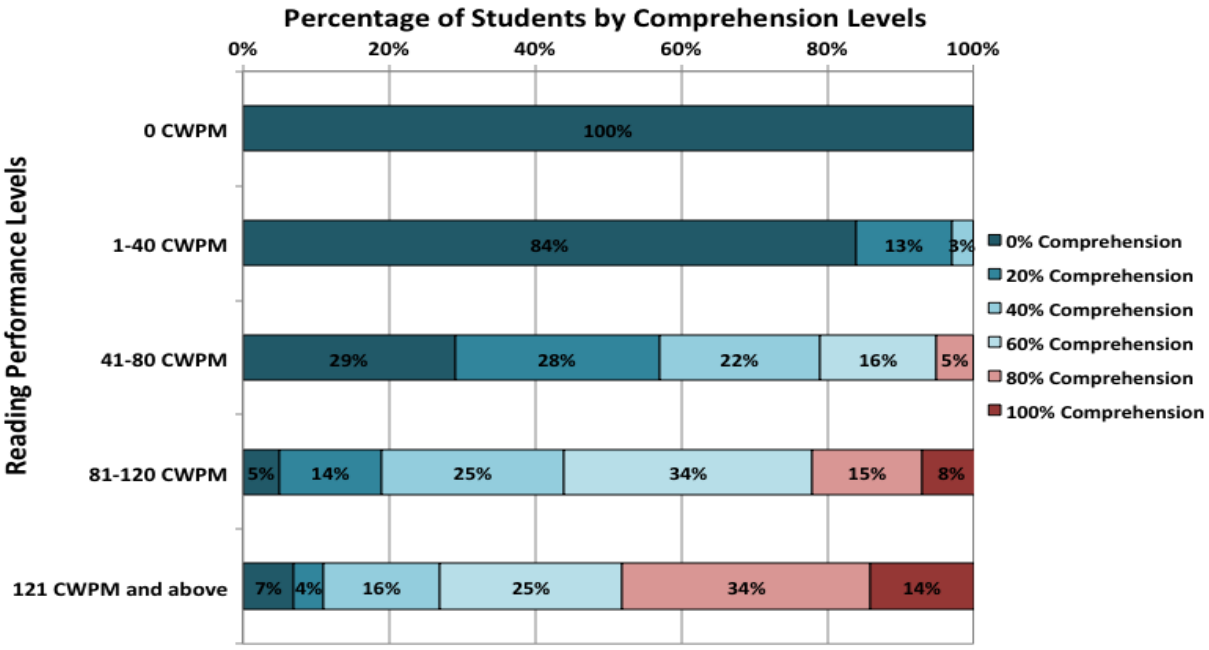
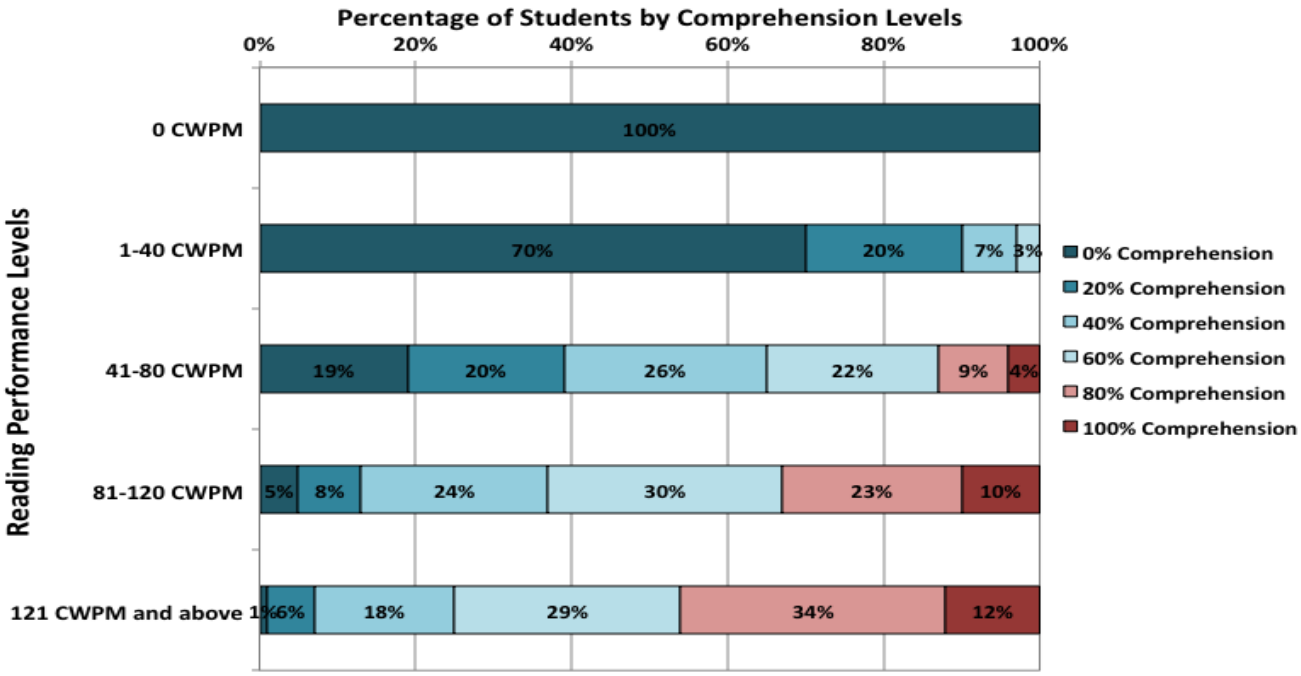


TABLE A5: GRADE 5 READING FLUENCY AND COMPREHENSION

Category (Performance Level)	WCPM	% of Students by Comprehension Level						
		0%	20%	40%	60%	80%	100%	Total
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	70%	20%	7%	3%	0%	0%	100%
Intermediate Reader	41 to 80	19%	20%	26%	22%	9%	4%	100%
Fluent Reader	81 to 120	5%	8%	24%	30%	23%	10%	100%
Advanced Reader	121 and above	1%	6%	18%	29%	34%	12%	100%

FIGURE A5: GRADE 5 READING FLUENCY AND COMPREHENSION



The main results for the categories of reading speed (from non-readers to advanced readers) in relation to comprehension levels (from 0 percent to 100 percent) for grades 3 and 5 are summarized as follows:

- Non-Readers (0 WCPM) – All of the non-readers had 0 percent comprehension.
- Early Readers (1-40 WCPM) – Most of the early readers (84 percent at grade 3 and 70 percent at grade 5) had 0 percent comprehension. None of them achieved 80 percent comprehension.
- Intermediate Readers (41-80 WCPM) – About one-fourth the intermediate readers (29 percent at grade 3 and 19 percent at grade 5) had 0 percent comprehension. A small minority of them (7 percent at grade 3 and 13 percent at grade 5) achieved at least 80 percent comprehension.
- Fluent Readers (81-120 WCPM) – About 5 percent of the fluent readers (at each grade level) had 0 percent comprehension. Less than one-third of them (23 percent at grade 3 and 33 percent at grade 5) achieved at least 80 percent comprehension.
- Advanced Readers (121 WCPM and above) – A small percentage of the advanced readers had 0 percent comprehension. Fewer than half of them (48 percent at grade 3 and 46 percent at grade 5) achieved at least 80 percent comprehension.

The key point from the data is that most of the fluent and advanced readers – at both grade levels – did not reach 80 percent comprehension. Setting a threshold under the assumption that fluent readers (in terms of speed) have a high level of comprehension can be misleading. Conversely, using a single indicator, i.e., reading speed, to set thresholds can be a more reliable way of interpreting the results.